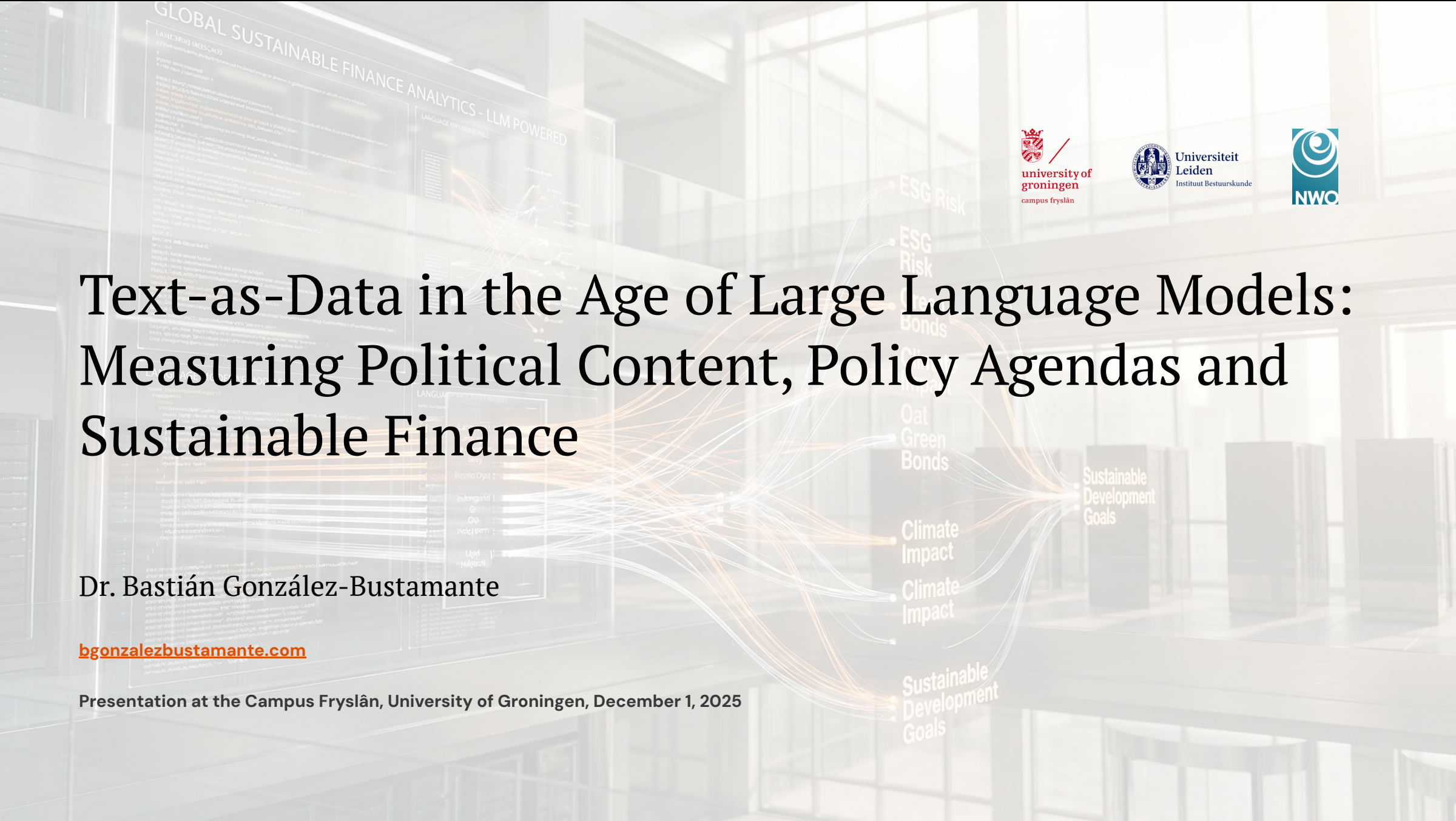


university of groningen
campus fryslân

Universiteit Leiden
Instituut Bestuurskunde

NWO

Text-as-Data in the Age of Large Language Models: Measuring Political Content, Policy Agendas and Sustainable Finance

Dr. Bastián González-Bustamante

bgonzalezbustamante.com

Presentation at the Campus Fryslân, University of Groningen, December 1, 2025

Transforming Text into Insight

Text is everywhere, yet **unlocking its value** for social science and policy research remains a significant challenge. By treating text-as-data, we can uncover profound insights across critical domains.

Political Conflict

Regulation and Policy

Social Toxicity

Sustainability

Financial Governance

Policy agendas

How can we turn messy text into reliable measurements for theory-driven social science and policy-relevant research?

university of groningen
campus fryslân

Presentation Roadmap

Evolution of Methods

We will trace the development of text analysis techniques, highlighting key advancements that allow us to extract deeper insights from unstructured data.

The VIDI Project

An in-depth look at our sustainable finance project, utilising advanced techniques from topic models to cutting-edge LLM-based RAG and multi-agent pipelines.

1

2

Real-World Applications

Discover how these methodologies are applied to critical areas such as understanding **online toxicity** and shaping **policy agendas**.

3

Evolution of Methods

From Classic Text-as-Data to LLMs

university of groningen
campus fryslân

The Challenge: Too Many Texts

The Problem

Language is central to social sciences, but scholars struggle with volume. Manually reading all texts in even moderately sized corpora is impossible. Hiring coders remains very expensive.

The Solution

Automated content methods make possible the previously impossible: systematic analysis of large-scale text collections without massive funding support.

The Classic Text-as-Data Toolkit

Bag-of-Words and Document-Term Matrices

Represent text as quantifiable data by focusing on word frequencies and co-occurrences, forming the basis for many analytical tasks.

Dictionaries and Lexicons

Leverage predefined word lists to measure specific concepts, sentiments, or themes in text, offering theory-driven insights.

Supervised Classification

Train models on labelled data to automatically categorize new texts, identifying specific patterns and assigning them to predefined classes.

Unsupervised Topic Models (LDA, STM)

Discover latent thematic structures within large text collections without prior labeling, revealing the main subjects and their distribution.

university of groningen
campus fryslân

Four Principles of Automated Text Analysis

All Models Are Wrong

But some are useful. Complexity of language means all methods fail to provide accurate accounts of data-generating processes.

Augment, Do not Replace

Quantitative methods amplify human abilities. Computers enhance, not eliminate, the need for careful thought and close reading.

No Best Method

Different research questions imply different models. Performance varies substantially across contexts and data sets.

Validate, Validate, Validate

Output may be misleading or wrong. Researchers must demonstrate reliability through experimental, substantive, and statistical evidence.

From Words to Numbers

Language is complex, but not all complexity is necessary. We transform text into quantitative data through preprocessing steps that retain useful information whilst discarding ancillary complexity.

1

Bag of Words

Discard word order. Simple word lists often suffice to convey general meaning.

2

Stemming

Reduce words to roots (e.g., family, families, familial → famili) to decrease vocabulary size.

3

Remove Noise

Discard punctuation, capitalisation, very common words (stopwords), and very rare words.

4

Document-Term Matrix

Each document becomes a vector counting word occurrences. Typically 3,000–5,000 features.

Dictionary Methods

The most intuitive automated method. Dictionaries use the rate at which key words appear to classify documents or measure concepts like tone.

Each word receives a score (e.g., positive = +1, negative = -1). Document tone is calculated as the weighted average of word scores.

Off-the-shelf dictionaries fail for corporate earnings reports. Words like "tax", "cost", "crude", or "cancer" have negative connotations elsewhere but may be positive in earnings reports.

Domain Specificity

Create problem-specific dictionaries or validate existing ones for your context.

Validation Challenge

Humans cannot reliably produce granular scales, making gold-standard evaluations essentially impossible.

university of groningen
campus fryslân

Supervised Learning Methods

A domain-specific alternative to dictionaries. Human coders classify a subset of documents (training set). The algorithm learns from this, then classifies remaining documents.

1

Construct Training Set

Develop coding scheme iteratively. Sample documents representatively.

2

Train Model

Algorithm learns relationship between features and categories.

3

Validate and Apply

Compare machine to human coding. Apply to test set.

Constructing a Training Set

Creating Coding Schemes

- Write concise codebook to guide coders
- Apply to initial documents
- Identify ambiguities and revise
- Iterate until no ambiguities remain (inter-coder)

Selecting Documents

- Training sets should be representative
- Use random or stratified sampling
- Rule of thumb: 500 documents (better 1,000)
- More categories require more documents

university of groningen
campus fryslân

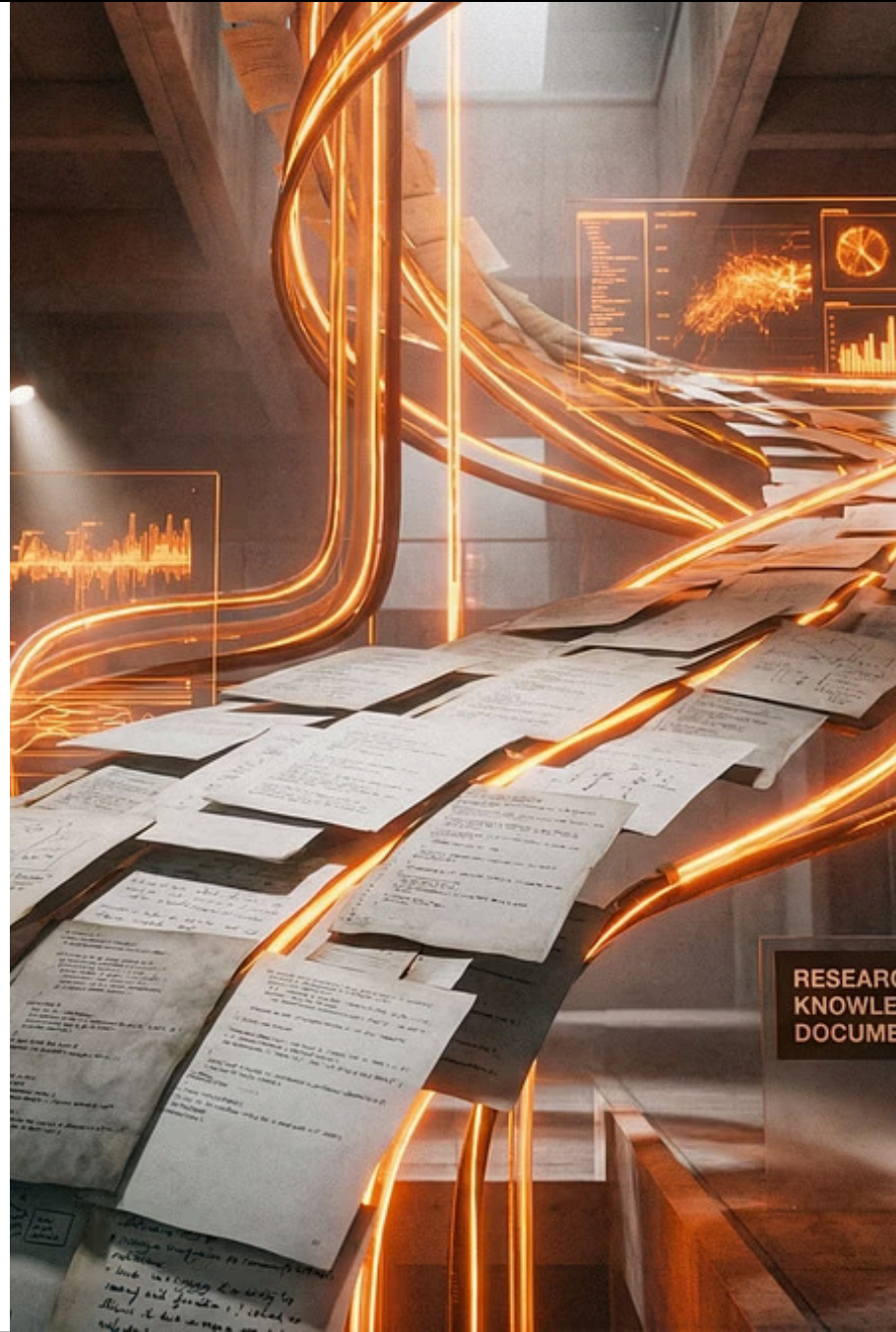
Naïve Bayes Classifier

A canonical supervised learning example. Uses Bayes's rule to infer document category given word profile.

$$p(C_k|W_i) \propto p(C_k)p(W_i|C_k)$$

The "naïve" assumption: given a document's category, words are generated independently. Though wrong, this model captures useful information for classification.

Other methods include **Random Forest**, **Support Vector Machines (SVM)**, and neural networks. Ensembles combine classifiers to improve accuracy and stability.



Topic Models: LDA

Latent Dirichlet Allocation (LDA) is the first and most widely used topic model. Topics are probability distributions over words. Documents are mixtures of topics.

Dirichlet Prior

Shares information across units

1

Mixed Membership

Documents allocate attention to topics

2

Word Assignments

Each word assigned to one topic

3

Scaling Political Actors

Locating political actors in ideological space using text. Nearly all political actors speak, making text-based scaling widely applicable.

Wordscores (Supervised)

Uses reference texts to define political positions. Generates word scores measuring separation. Scales remaining texts using weighted averages.

Wordfish (Unsupervised)

Discovers words distinguishing locations. Poisson-IRT model assumes politician's position affects word usage rates.

university of groningen
campus fryslân

The LLM Revolution: Beyond Traditional Text-as-Data

Transfer Learning

Models pre-trained on massive datasets can be fine-tuned for specific tasks with minimal new data.

Instruction Following

LLMs can interpret and execute complex instructions, moving beyond simple classification to sophisticated task completion.

Multilinguality

Many LLMs inherently support multiple languages, enabling cross-lingual analysis without separate models.

Zero/Few-Shot Learning

Perform tasks with few or no examples, drastically reducing the need for extensive training data.

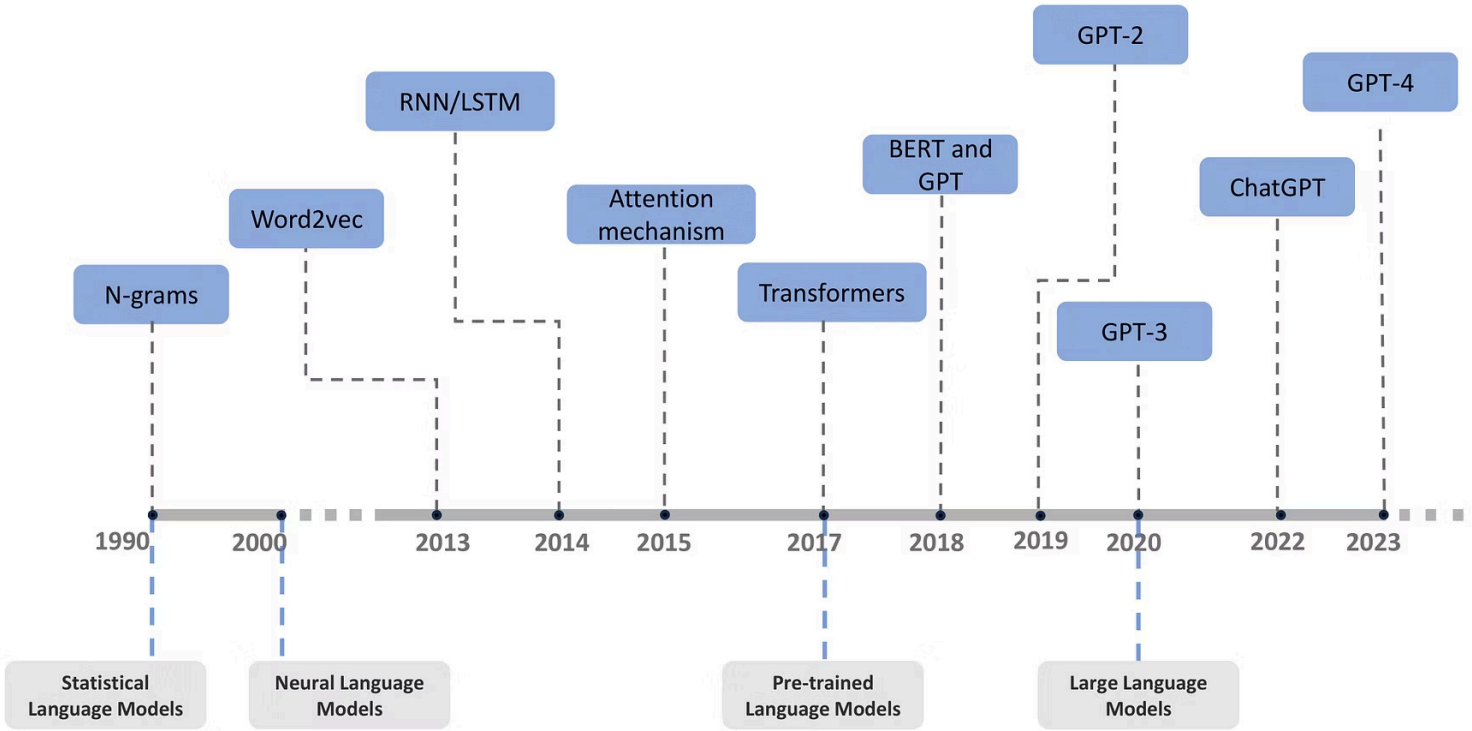
Text Generation

Generate summaries, explanations, or synthetic data, opening new avenues for research and applications.

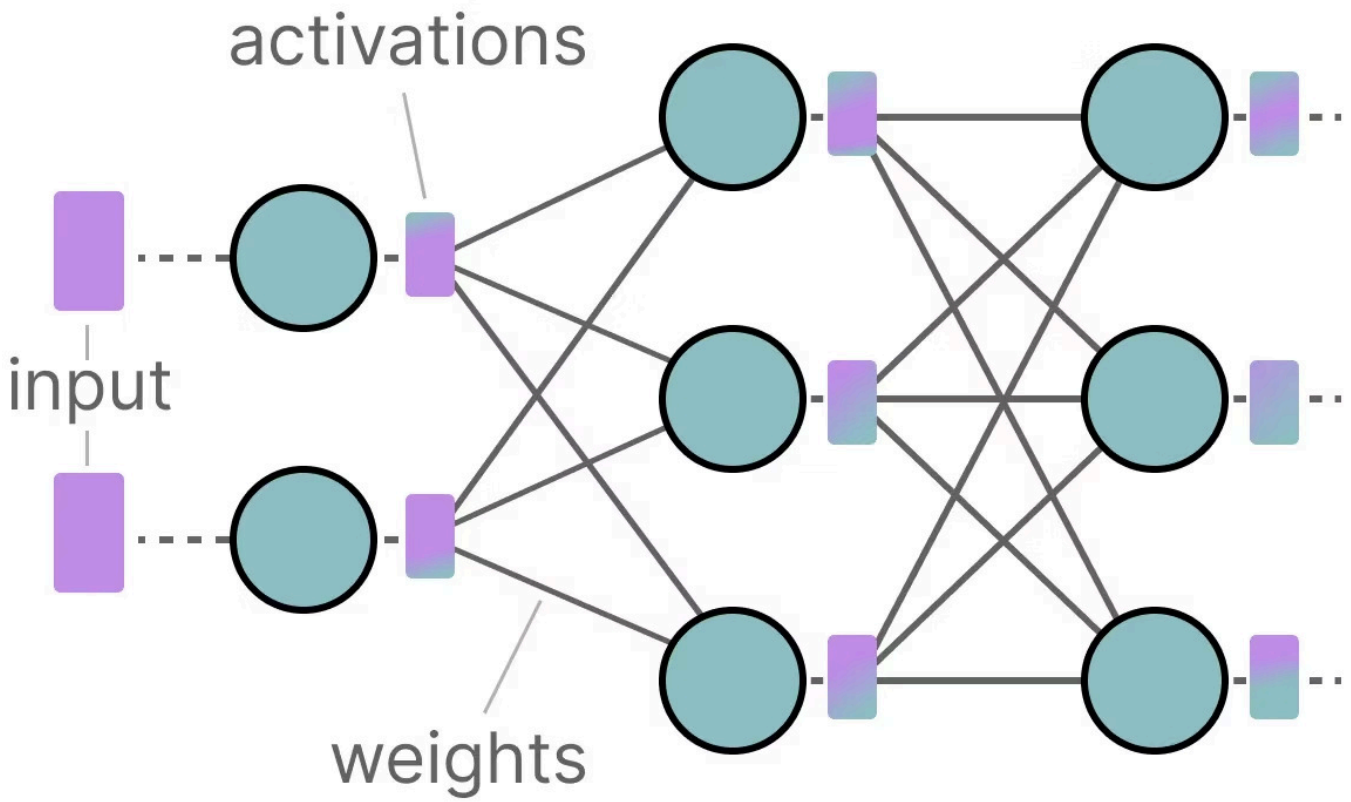
Reasoning Pipelines

LLMs can be orchestrated into complex workflows to perform multi-step reasoning and problem-solving.

Evolution of LLMs

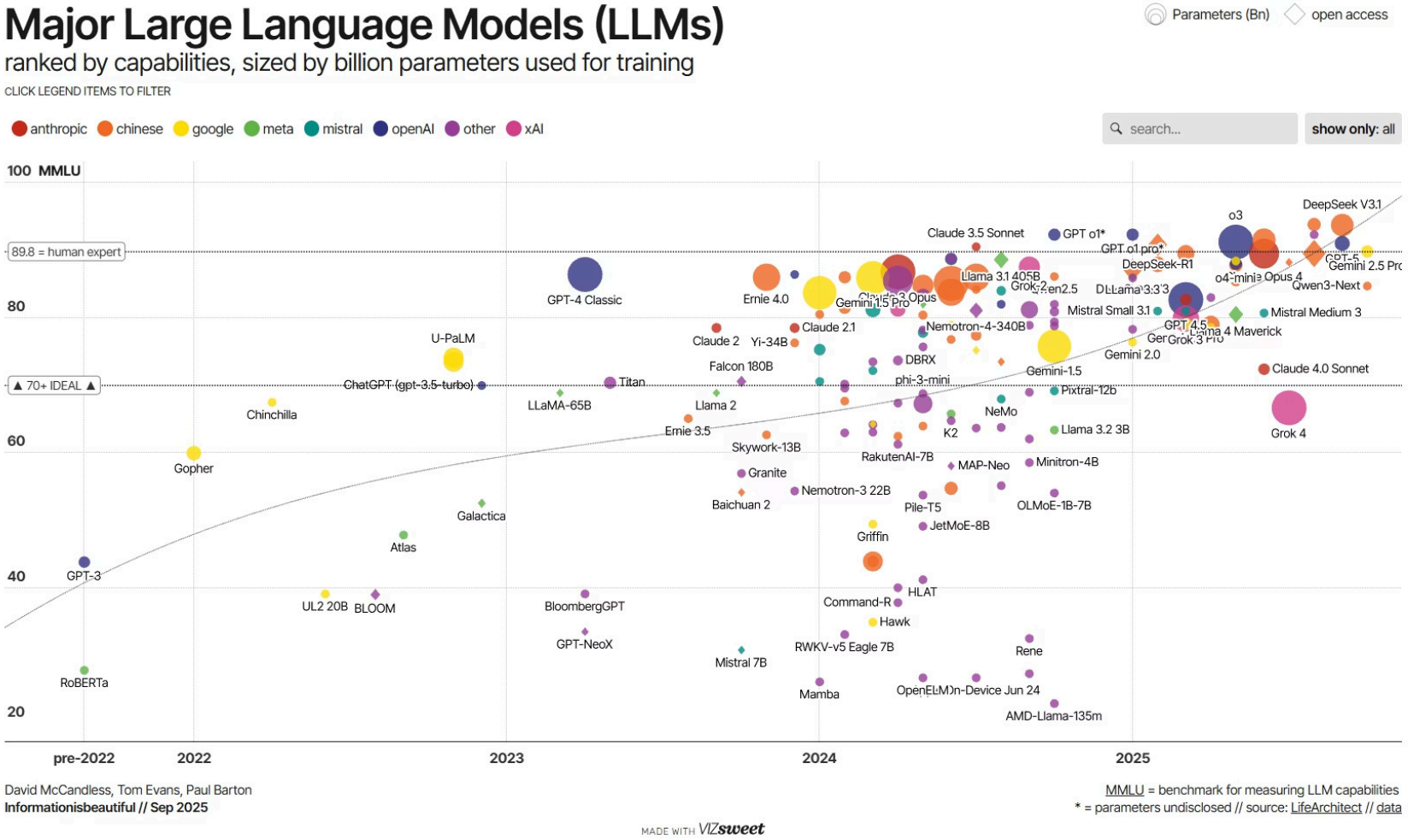


Neural Networks



Major Large Language Models (LLMs)

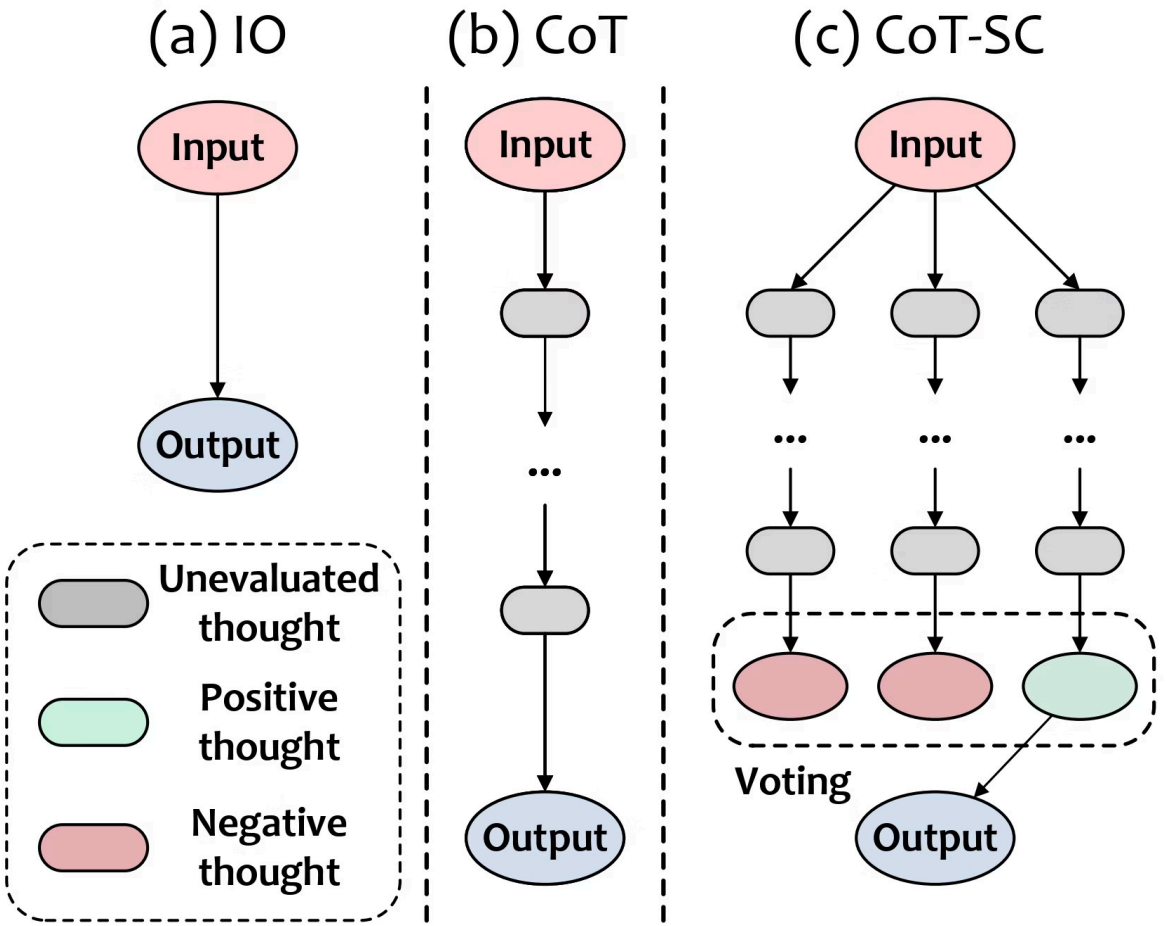
ranked by capabilities, sized by billion parameters used for training



David McCandless, Tom Evans, Paul Barton
Informationisbeautiful // Sep 2025

MMLU = benchmark for measuring LLM capabilities
* = parameters undisclosed // source: LifeArchitect // data

MADE WITH VIZsweet


(a) IO
Input
Output
Unevaluated thought
Positive thought
Negative thought
(b) CoT
Input
...
Output
(c) CoT-SC
Input
...
Voting
Output

Real-World Applications

Case 1: Online Toxicity

university of groningen
campus fryslân

LLMs in Multilingual Toxic Speech Detection

Online Toxicity Crisis

Hate speech, harassment, and abusive language undermine digital discourse, depressing participation and eroding trust in public deliberation.

Anonymity and global reach of social media intensify these effects, demanding sophisticated detection systems.

The Detection Problem

What one culture judges as banter, another may deem a slur. Context, speaker identity, and power asymmetries complicate judgements.

Universal rule-sets prove impractical. We need nuanced, contextual reasoning.

university of groningen
campus fryslân

Our Approach

Comprehensive Benchmark

Tested **807 model-language pairs** across seven languages using the TextDetox corpus.

State-of-the-Art Models

Evaluated OpenAI's GPTs, o-series, Claude, Grok, Llama, Qwen, Mistral, and others.

Strict Zero-Shot

Applied in-context learning with deterministic and default temperature runs.

Meta-Analysis

Pooled goodness-of-prediction indicators to identify patterns across languages and models.

university of groningen
campus fryslân

Languages Tested

High-Resource Languages

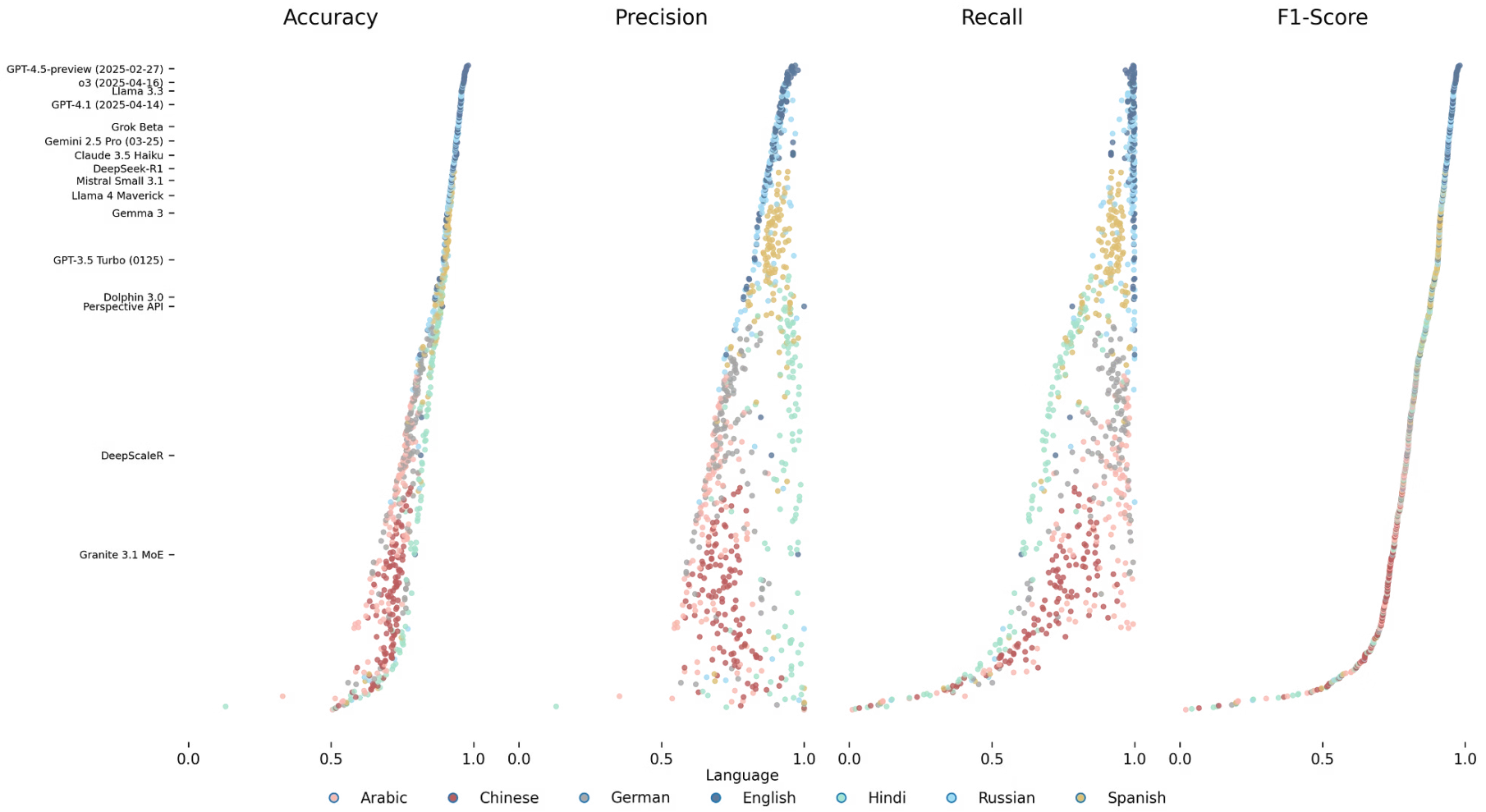
- English
- German
- Spanish

Abundant training data and well-curated corpora available.

Low-Resource Languages

- Arabic
- Chinese
- Hindi
- Russian

Scarcer annotated resources and wider dialectical variation.


Accuracy
Precision
Recall
F1-Score
GPT-4.5-preview (2025-02-27)
o3 (2025-04-16)
Llama 3.3
GPT-4.1 (2025-04-14)
Grok Beta
Gemini 2.5 Pro (03-25)
Claude 3.5 Haiku
DeepSeek-R1
Mistral Small 3.1
Llama 4 Maverick
Gemma 3
GPT-3.5 Turbo (0125)
Dolphin 3.0
Perspective API
DeepScaleR
Granite 3.1 MoE
0.0
0.5
1.0
Language
Arabic
Chinese
German
English
Hindi
Russian
Spanish

Takeaways

+7.7%

Resource Advantage

High-resource languages enjoy a 7.7 points F1 advantage

-9.8%

Size Penalty

Small models (≤7B) fall 9.8 points short of larger LLMs baseline

university of groningen
campus fryslân

The Overthinking Penalty

Why Reasoning Does not Help

Toxicity Detection Requires Simplicity

This task demands **binary judgement**, necessitating **rapid, context-sensitive decisions** rather than elaborate, deep reasoning processes.

Reasoning Can Hinder Accuracy

Chain-of-thought prompts often decrease accuracy by leading models into irrelevant moral digressions or hallucinating context not present in the text.

CoT's Limited Impact on Classification

While **chain-of-thought** performs better on maths or logic tasks, its gains on classification are notably smaller, and can even be detrimental in specific contexts like toxicity detection.

university of groningen
campus fryslân

Evidence-Based Recommendations

Use Open Models for High-Resource Languages

English, German, Spanish: open-source models like Granite 3.2 or Nous-Hermes 2 Mixtral are competitive without performance penalty.

Choose Proprietary for Low-Resource Languages

Arabic, Chinese, Hindi, Russian: closed models like GPTs or Claude justify API costs with superior performance.

Skip Reasoning for Binary Tasks

Do not pay extra for reasoning models on toxicity detection. Standard LLMs are more suitable and cost-effective.

Small Models as Edge Solutions Only

Use compact models for prototyping or when memory and speed are critical, but expect 10-point performance penalty.

Deployment Guidelines

Fix Temperature

Set to zero for full reproducibility. If unavailable, higher temperatures yield nearly identical annotations.

Continuous Evaluation

Rapid model releases mean today's evaluation can change within months. Maintain lightweight test suites.

Tiered Strategy

Open models for high-resource languages, closed or hybrid for low-resource languages.

Real-World Applications

Case 2: Policy Agendas

university of groningen
campus fryslân

A Comprehensive Benchmark of LLMs for Multilingual Policy Agenda Topic Annotation

We present a systematic benchmark of **79 state-of-the-art LLMs** on a substantively important and methodologically challenging task: annotating major policy topics from the Comparative Agendas Project.

Multilingual Corpus

Testing across Danish, Dutch, English, French, Hungarian, Italian, Portuguese and Spanish legislative texts.

Comprehensive Models

Comparing proprietary models (GPTs, o-series, Claude, Grok) and open-source alternatives (Llama, Qwen, Mistral).

Reasoning Assessment

Evaluating whether models optimised for chain-of-thought reasoning outperform general-purpose LLMs.

university of groningen
campus fryslân

The Comparative Agendas Project

A Common Framework

CAP provides a hierarchical codebook allocating policy-relevant texts to one of 21 major topics and over 200 subtopics, enabling cross-country comparison.

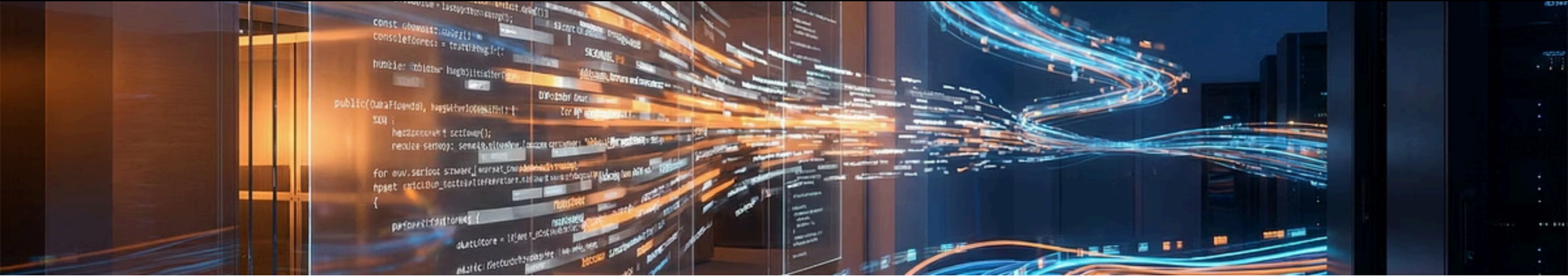
Political attention is a scarce resource. Governments can only handle a handful of problems simultaneously, making agenda-setting a competitive, path-dependent process.

21

Major Topics

200+

Subtopics



Methodological Challenges

Resource Intensity

Maintaining inter-coder reliability above 70–80% is time-consuming and expensive, especially for multilingual corpora requiring extensive training.

Automation Limitations

Early efforts using dictionaries and classic machine learning rarely exceeded F1-scores of 0.70 and generalised poorly across domains.

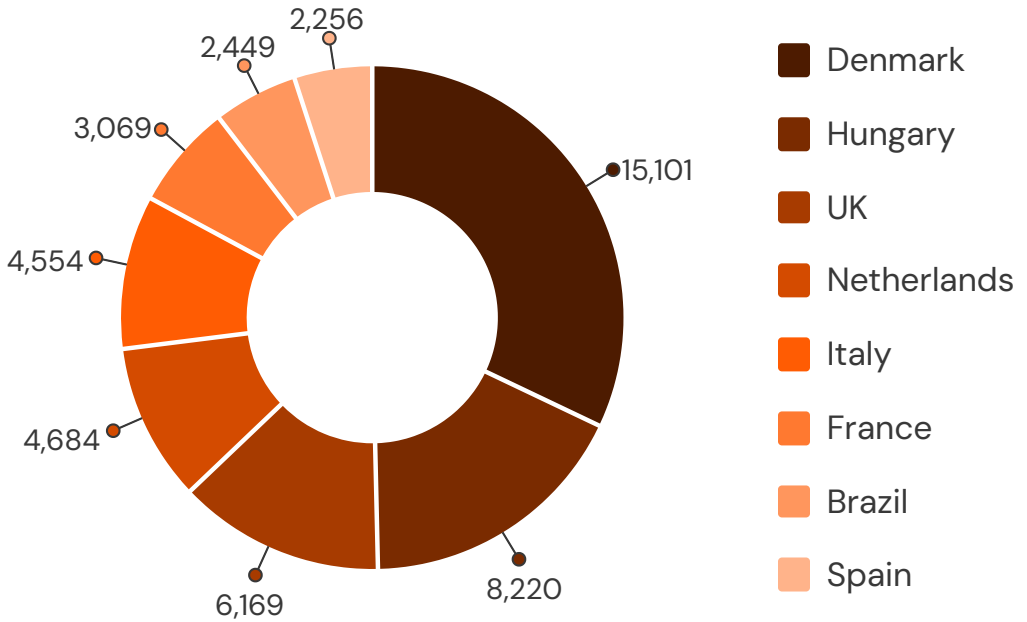
Conceptual Validity

CAP's single-label rule forces selection of one primary topic, potentially obscuring cross-cutting frames like environmental and energy legislation.

Ground-Truth Data

We assembled eight country corpora from the Comparative Agendas Project, spanning over a century of legislative activity across multiple languages.

Each corpus was stratified into 70% training, 15% validation and 15% test sets, maintaining relative topic prevalence across splits.



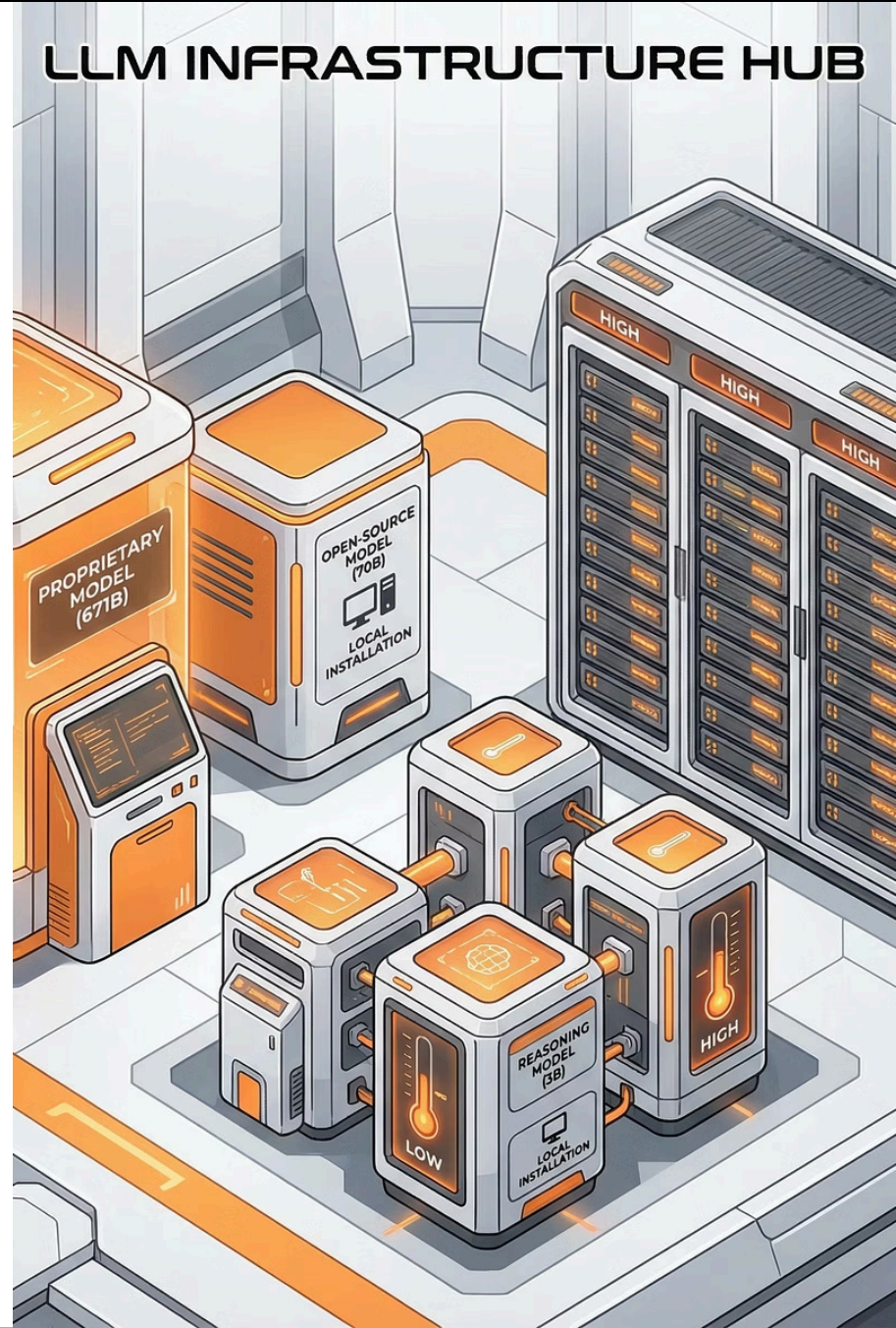
university of groningen
campus fryslân

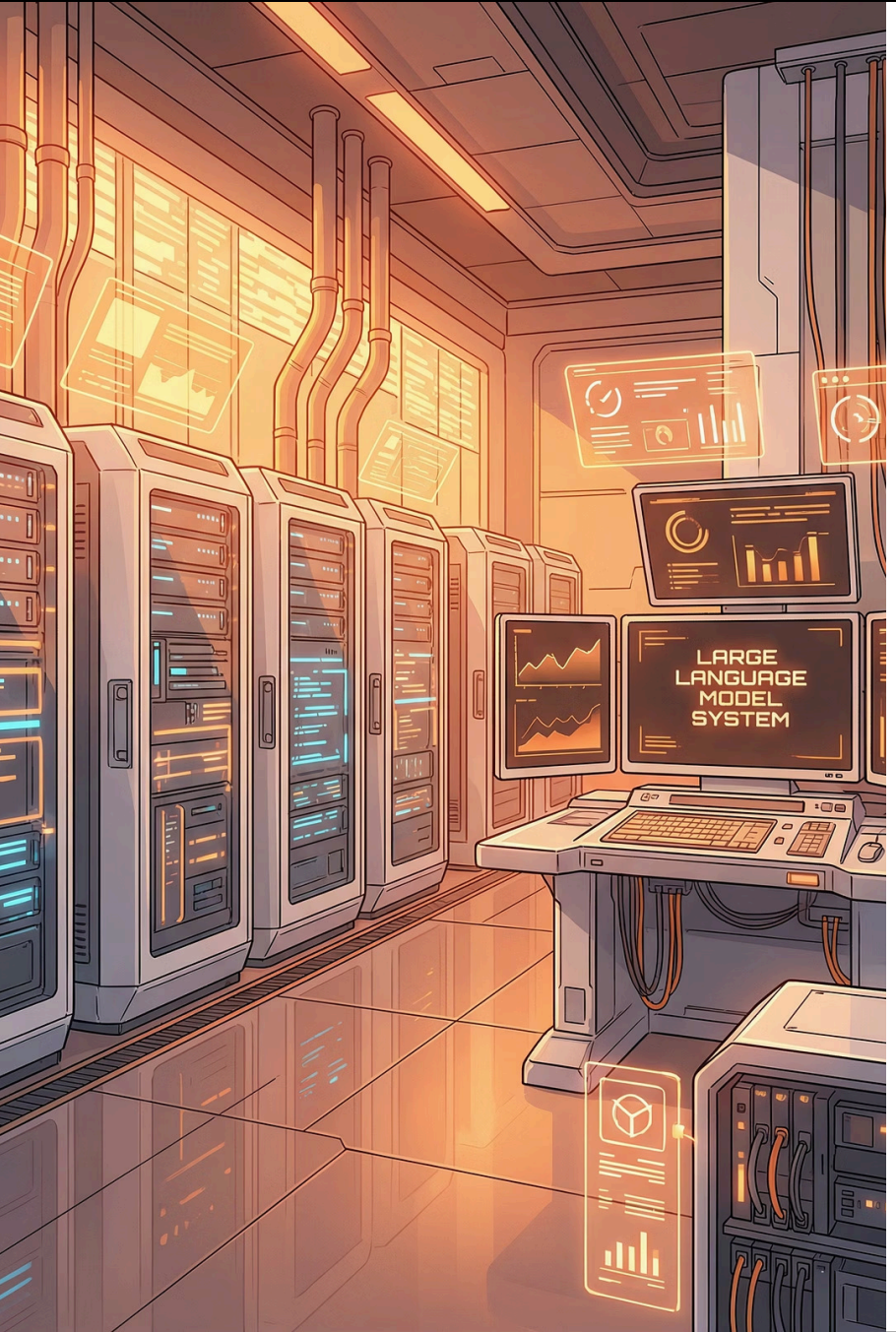
The LLM Landscape

79 Contemporary Models

- Proprietary reasoning models (OpenAI o-series)
- Proprietary general models (GPTs, Grok, Gemini)
- Open-source reasoning models (DeepSeek-R1, OpenThinker)
- Open-source general models (Llama, Qwen, Mistral)

Models ranged from 3 billion to 671 billion parameters, deployed via API or locally with controlled temperature settings.





university of groningen
campus fryslân

Zero-Shot Classification Protocol

1

Uniform Prompt

Identical English system prompt instructing assignment of one CAP major topic label to legislative text.

2

Pure Zero-Shot

21 label names listed with codebook definitions. No examples provided, ensuring genuine zero-shot evaluation.

3

Controlled Inference

Reasoning models at recommended temperature, open-source models at zero temperature for deterministic outputs.

university of groningen
campus fryslân

Meta-Analysis Framework

We conducted random-effects meta-regression on **620 model-language pairs**, analysing the logit-transformed F1-score to separate sampling error from between-study heterogeneity.

Baseline Model

Binary variable for reasoning LLMs to establish fundamental performance differences.

Source Effects

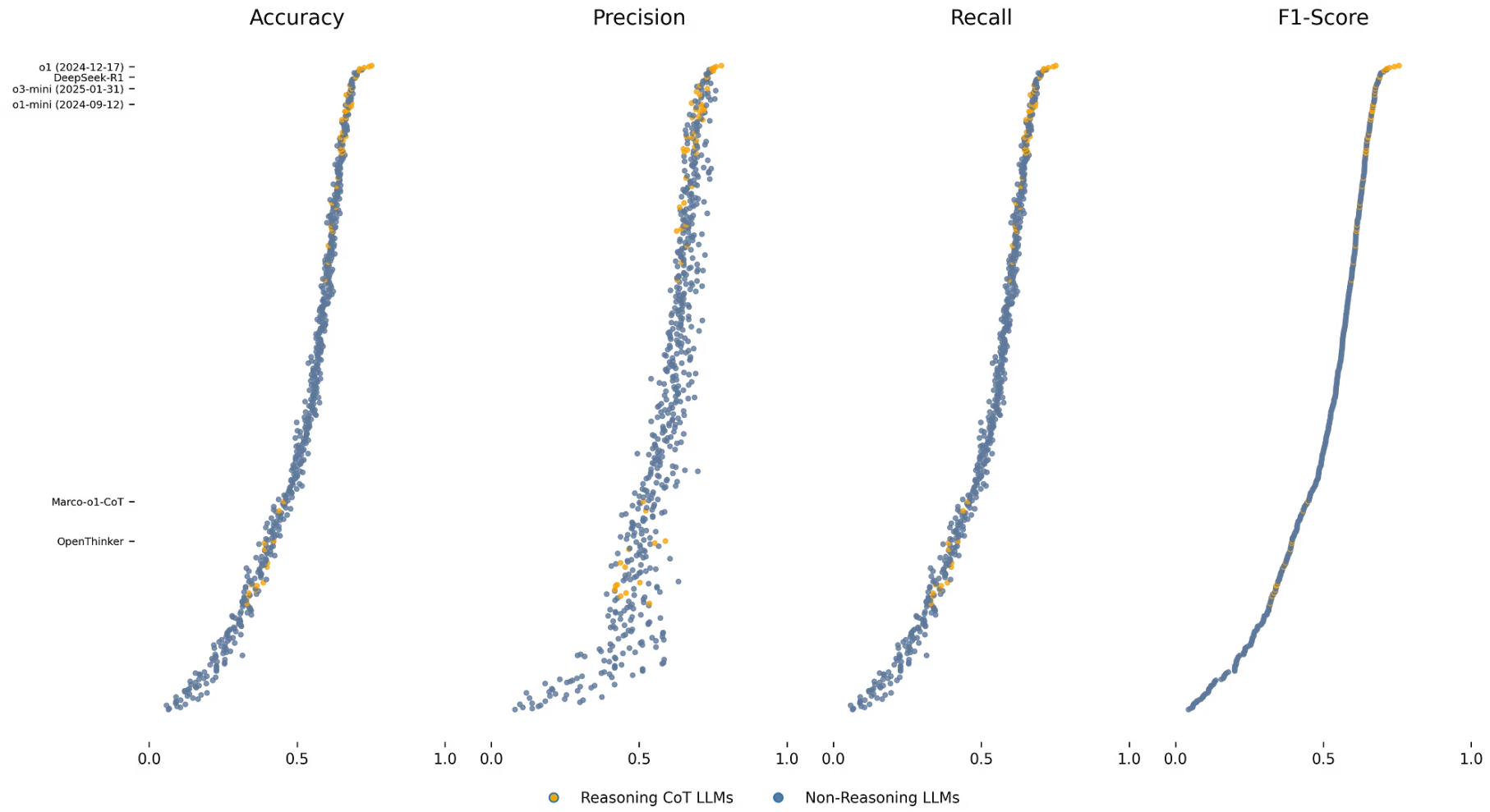
Incorporated open-source binary variable to assess proprietary versus open model performance.

Reproducibility Trade-offs

Added deterministic setup variable reflecting local deployment at temperature zero.

Full Specification

Controlled for parameters and language fixed effects using English as reference category.


Accuracy
Precision
Recall
F1-Score
o1 (2024-12-17)
DeepSeek-R1
o3-mini (2025-01-31)
o1-mini (2024-09-12)
Marco-o1-CoT
OpenThinker
0.0
0.5
1.0
Reasoning CoT LLMs
Non-Reasoning LLMs

Overall Performance Patterns

Top Performers

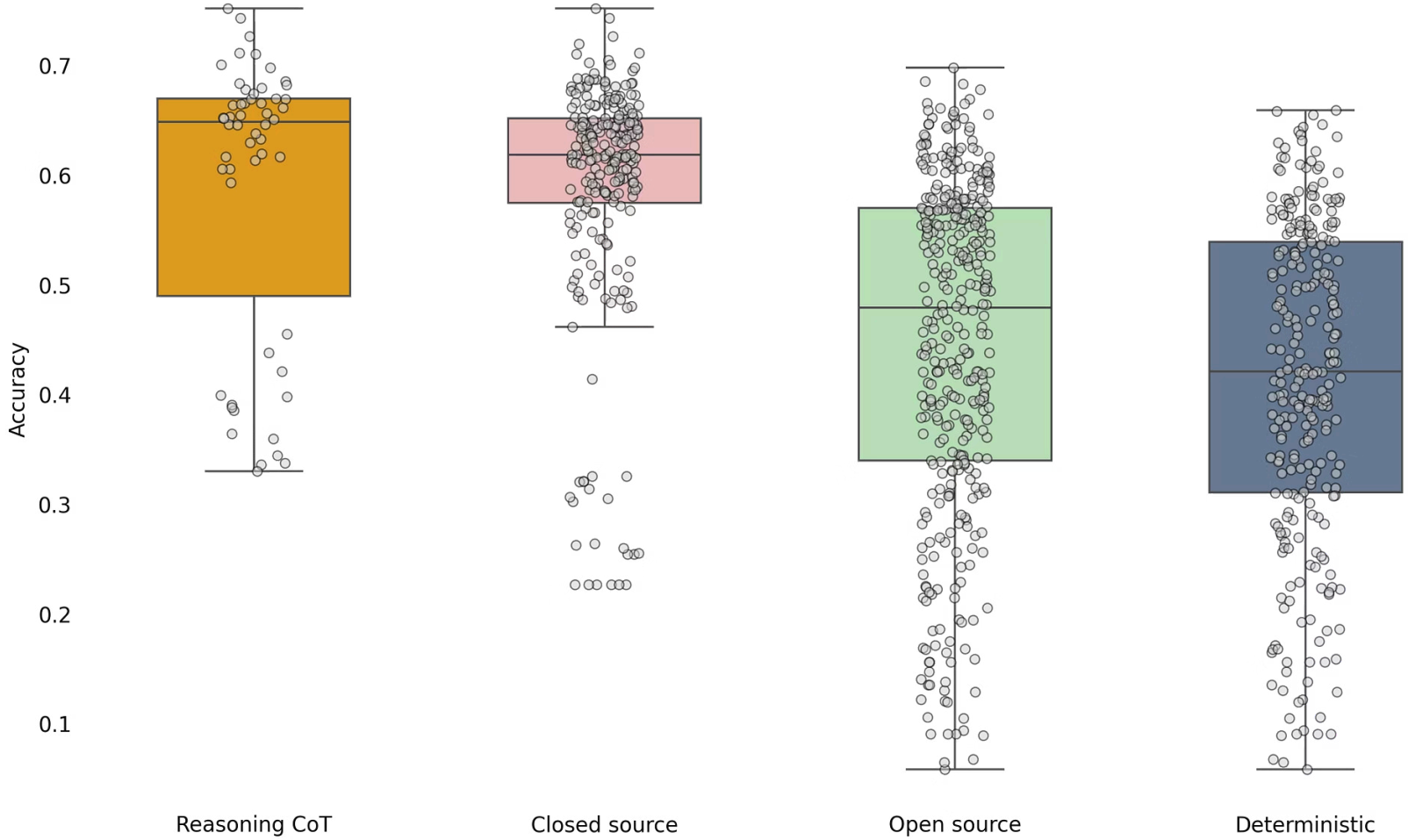
OpenAI **reasoning models** and DeepSeek-R1 achieved F1-scores above 0.60 across all languages, with **maximum scores reaching 0.76**.

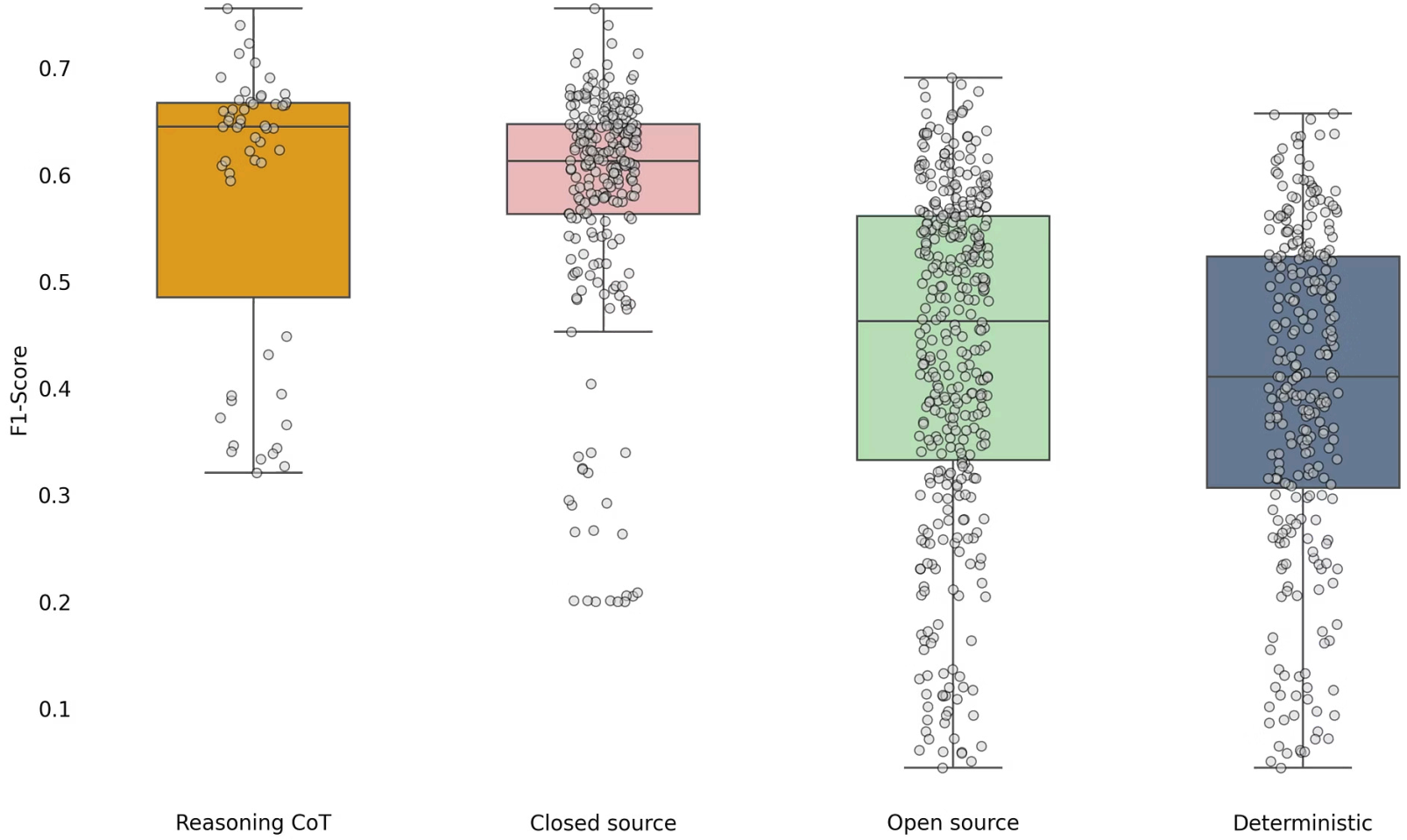
Middle Band

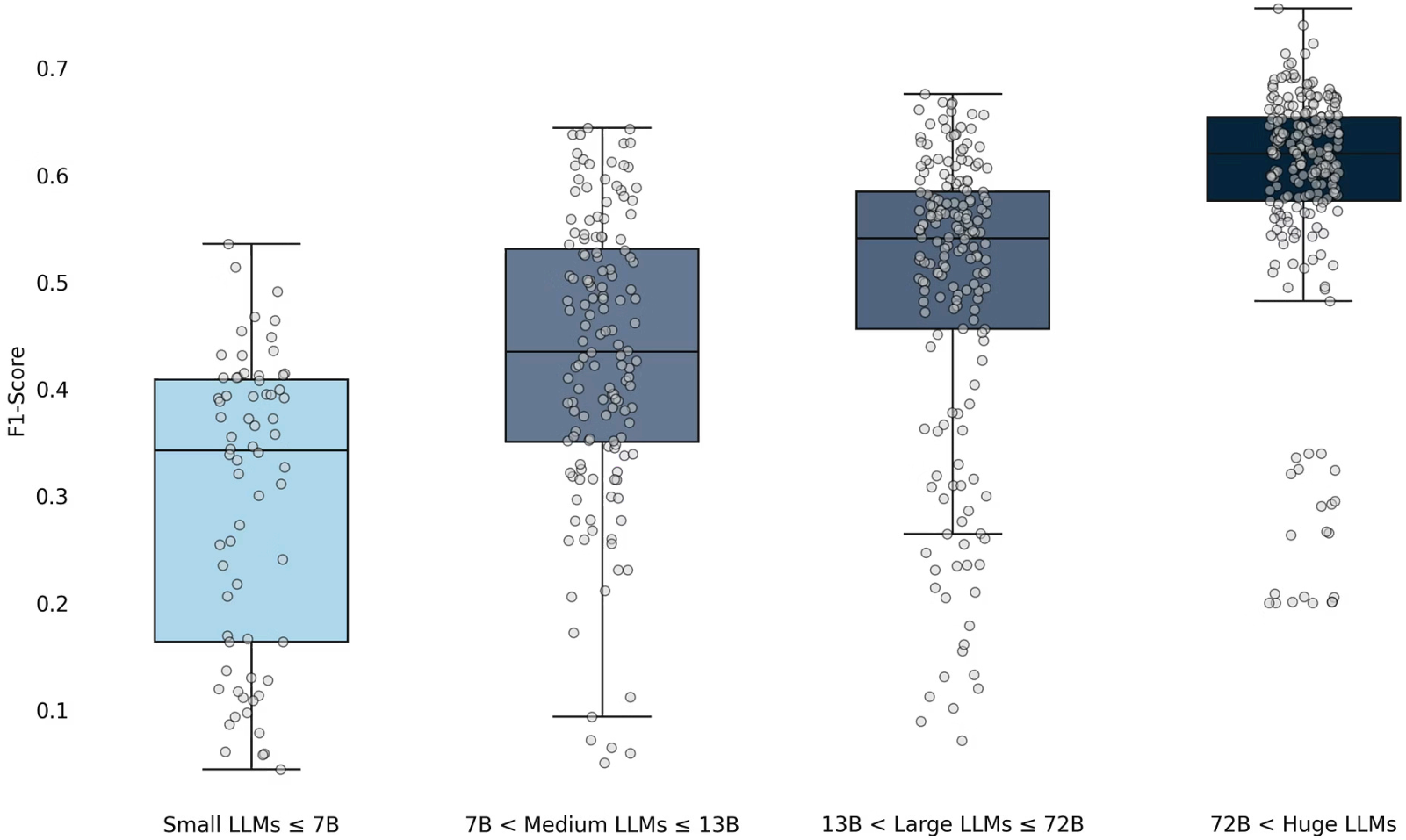
Closed models like Grok 3 and Gemini 2.0+ occupied **F1-scores between 0.58 and 0.70**, outperforming most open-source alternatives.

Long Tail

Open-source models displayed wide variation. Larger models approached proprietary performance, whilst **smaller models fell below 0.10 F1-score**.


Accuracy
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Reasoning CoT
Closed source
Open source
Deterministic


F1-Score
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Reasoning CoT
Closed source
Open source
Deterministic


F1-Score
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Small LLMs ≤ 7B
7B < Medium LLMs ≤ 13B
13B < Large LLMs ≤ 72B
72B < Huge LLMs

Takeaways

+8.9%

Reasoning Advantage

Chain-of-thought capabilities boost F1-score by almost 9 points

-7.8%

Reproducibility Penalty

Deterministic deployment reduces F1-score by about 8 points, but ensures consistency

university of groningen
campus fryslân

Fine-Tuning: The Best Option

In English, fine-tuned BERTs achieve **~10 improvement over best zero-shot LLM**, suggesting task-specific tuning can close the reasoning advantage gap.

Fine-tuned XLM-RoBERTa

Fine-tuned ModernBERT

The VIDI Project

Making Finance Sustainable?

A Comparative Perspective on the Politics of Investment

Prof. dr. Natascha van der Zwan, Principal Investigator

The Climate Finance Challenge

Investment Gap

The climate transition requires an estimated **$8.4 trillion** in annual investments until 2030, totalling **$266 trillion by 2050**. Current global sustainable investments by institutional investors stand at approximately $30 trillion.

Regional Variation

Industry reports demonstrate that some countries exhibit significantly higher levels of sustainable investments than others, suggesting certain political economies may be more successful at producing sustainable investors.

university of groningen
campus fryslân

The Challenge of Sustainable Finance Research

Data Scarcity

Comparative political economy research on sustainable finance faces a critical obstacle: the absence of consistent cross-national data. Traditional approaches struggle to capture the complexity of sustainability practices across different institutional contexts.

Narrative Complexity

Investors' annual reports contain rich sustainability narratives, but their unstructured nature makes systematic comparison difficult. The spectrum ranges from genuine commitments to greenwashing, complicating standardised measurement.

Research Gap and Objectives

The Problem

The correlation between sustainable investment outcomes and politico-economic characteristics remains unclear due to absence of systematic cross-country studies and lack of official SI data.

Our Approach

We examine magnitude and nature of sustainable investments amongst insurance companies, pension funds and sovereign wealth funds in relation to institutional variables.

The Innovation

Using text-as-data approaches and topic modelling to analyse annual reports, we uncover patterns in sustainability discourse across asset owners.

university of groningen
campus fryslân

Methodological Innovation

Text-as-data approaches in Natural Language Processing (NLP) offer solutions to methodological and empirical challenges facing sustainable finance studies. These techniques enable uncovering patterns, relationships and prevalence of particular topics by examining large volumes of text.

Document Acquisition

Collecting annual reports, sustainability reports and disclosure documents from asset owners.

Text Processing

Applying advanced extraction tools and translation methods to create a reliable corpus.

Topic Modelling

Using Structural Topic Models (STM) to identify common topics and measure their prevalence.

Statistical Analysis

Conducting regression analysis to evaluate effects of institutional and environmental variables.

Top 100 European Asset Owners

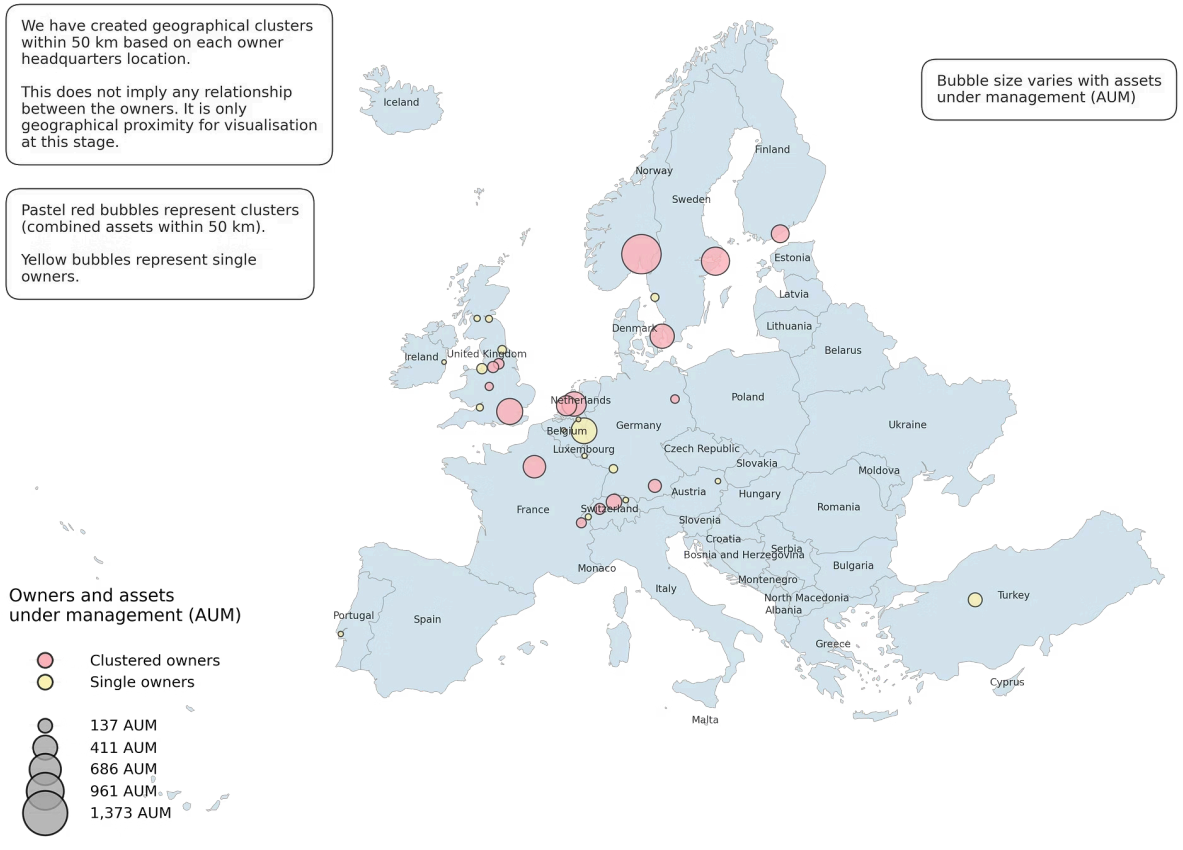
We have created geographical clusters within 50 km based on each owner headquarters location.

This does not imply any relationship between the owners. It is only geographical proximity for visualisation at this stage.

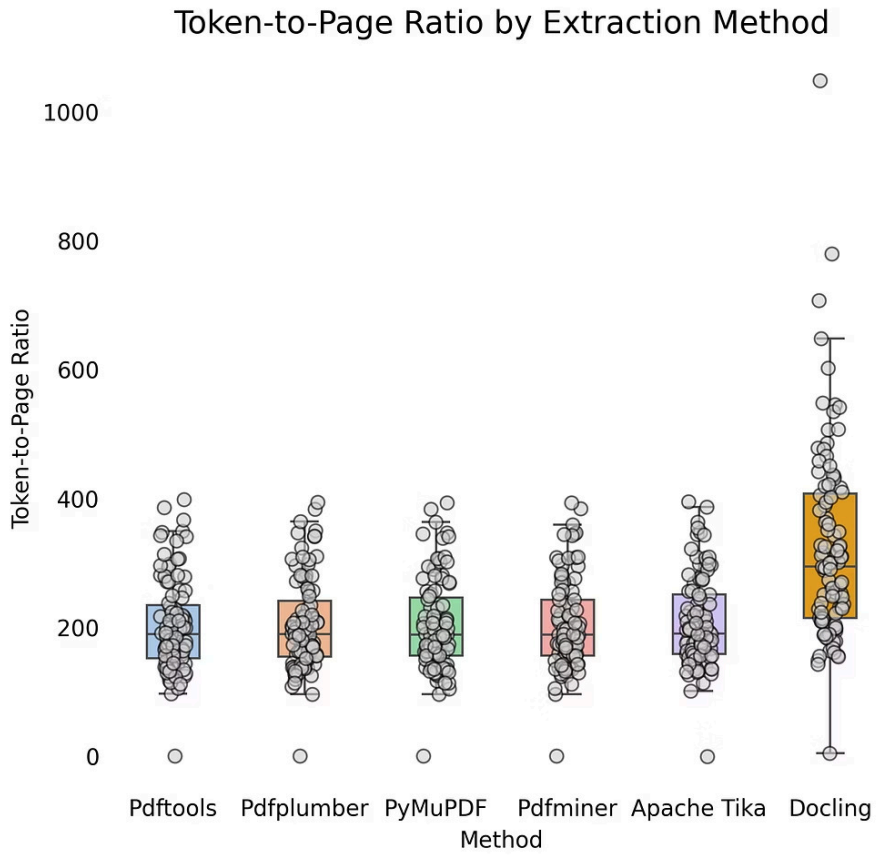
Pastel red bubbles represent clusters (combined assets within 50 km).

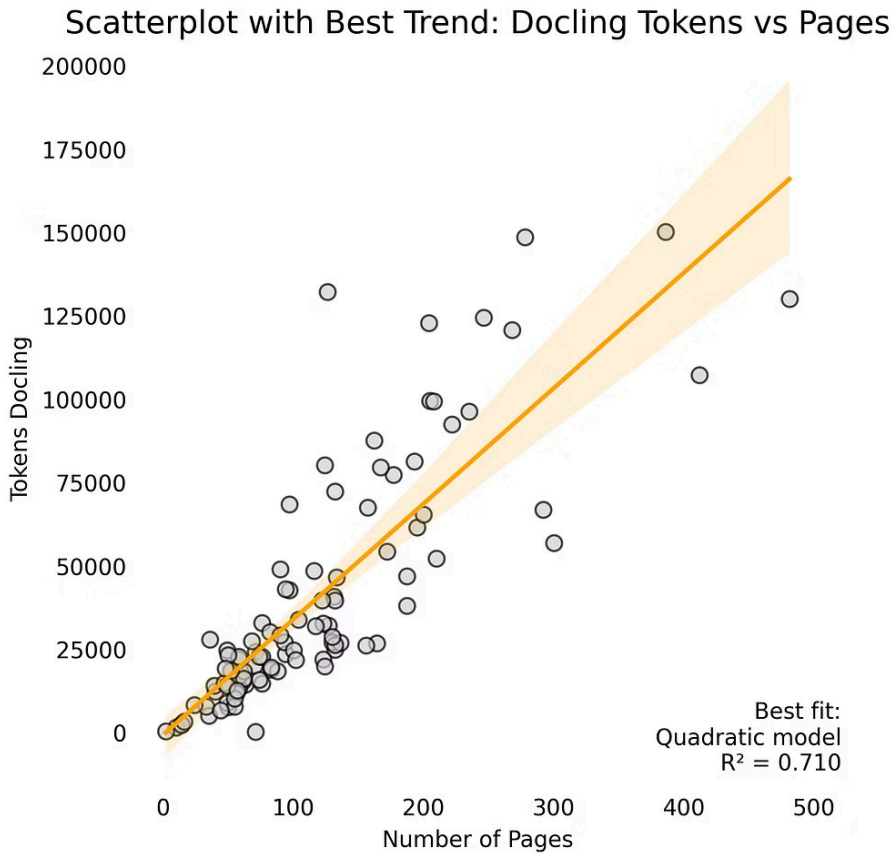
Yellow bubbles represent single owners.

Bubble size varies with assets under management (AUM)



Advanced Document Processing





university of groningen
campus fryslân

Structural Topic Modelling

We applied Structural Topic Modelling (STM) to systematically identify and analyse topic prevalence across annual reports. Unlike traditional approaches, **STM permits incorporation of document-level metadata**, making it particularly suitable for examining how institutional factors shape sustainable investment topics.

Why STM?

STM directly integrates contextual variables into the modelling process, allowing examination of how thematic content varies across different regulatory environments and languages.

Pre-Processing Pipeline

Standard NLP pipeline included converting text to lowercase, removing punctuation and numerals, applying tokenisation and discarding frequently used and rare tokens.

Topic Selection

Set STM to extract 15 topics driven by semantic coherence and exclusivity scores, ensuring topics are internally coherent and sufficiently distinct.

university of groningen
campus fryslân

Climate Emission Reporting	Occupational Pension Schemes and Member Befits	Pension Liabilities
Corporate Governance and Diversity	Contribution Monitoring and Regulatory Compliance	Reporting Standards and Strategy Assessments
Portfolio Management and Performance	Sustainability as Market Practice	Document Formatting and Linguistic Artefacts
Financial Operations	Group Assets and Credit Risk	Climate Leadership
Corporate and Government Bond Holdings	Pension Investment and Fund Returns	Pension Assets

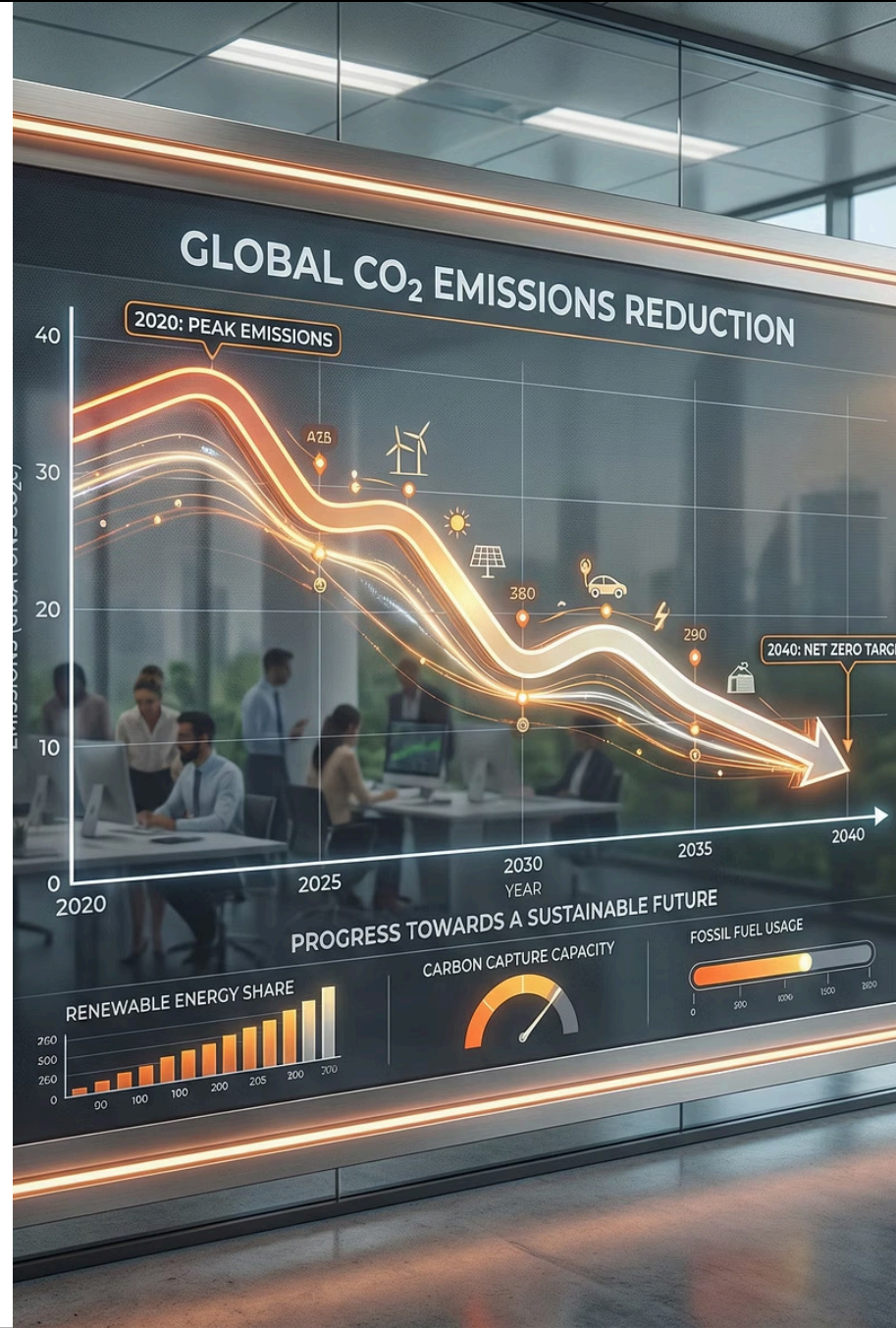
Topic: Climate Emissions Reporting

Dutch Pension Fund Focus

This topic emphasises **CO_2 emissions measurement**, featuring Dutch pension fund acronyms (PGGM, PME, BPL) with clear climate reduction targets and references to financed emissions.

> For 2023, the financed greenhouse gas emissions amount to 43.06 tCO_2 per million euros of invested capital. This represents a decrease of -38% compared to the reference year 2019.

These asset owners are amongst 61 financial sector organisations who signed the Netherlands' **Climate Commitment**, obligating annual reporting of investment portfolio CO_2 emissions.



Topic: Sustainability as Market Practice

This topic captures financial actors' attempts to **integrate sustainability with existing market practices**. Asset owners combine emphasis on sustainability with concerns for financial risks and returns, promoting new financial products marketed as sustainable investments.

> MN ensures positive impact by delivering a solid return with the right balance of risk for participants and retirees, as well as investments that contribute to a better world through impact investing and ESG integration.

Sustainability is mixed with financial considerations and reputational benefits, demonstrating how financial market practices shape environmental and governance strategies across business operations.

Topic: Climate Leadership

Swedish Asset Owners

This topic focuses on **Swedish pension funds and insurance companies'** climate commitments and goals, highlighting ambitions to achieve net zero emissions by 2050.

> As a major player and investor, we have the opportunity to exert influence and contribute to sustainable development. Our goal is to be the leader in the field of sustainability.

Swedish asset owners operate in a regulatory environment emphasising the **Climate Act** (net-zero by 2045) and the government's Fossil Free initiative.

university of groningen
campus fryslân

Econometric Analysis Strategy

Dependent Variables

Per-document topic relevance values (theta values) from STM for sustainability-related topics.

Key Predictors

Assets under management (quartiles), greenhouse gas emissions per capita, and carbon intensity of GDP.

Control Variables

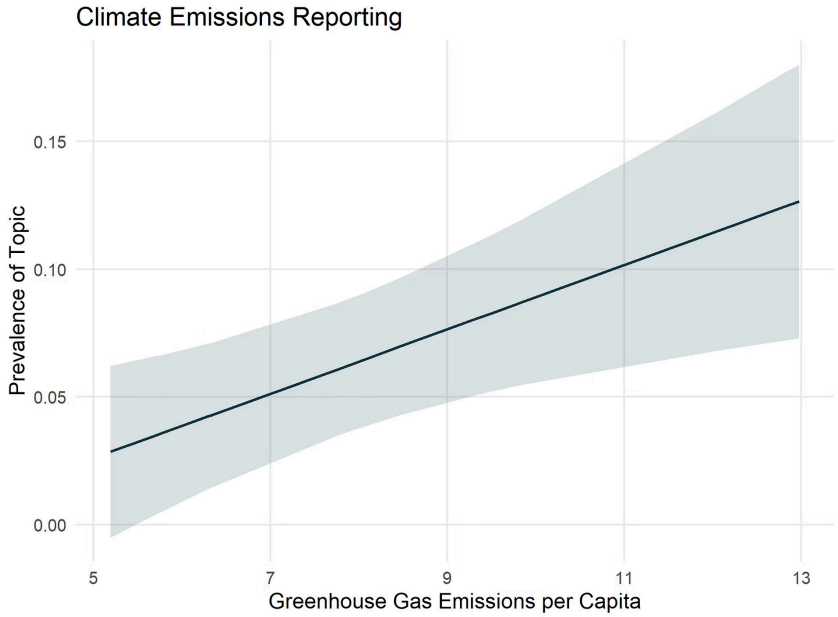
GDP growth and inflation to gauge the effect of general macroeconomic conditions on sustainability discourse.

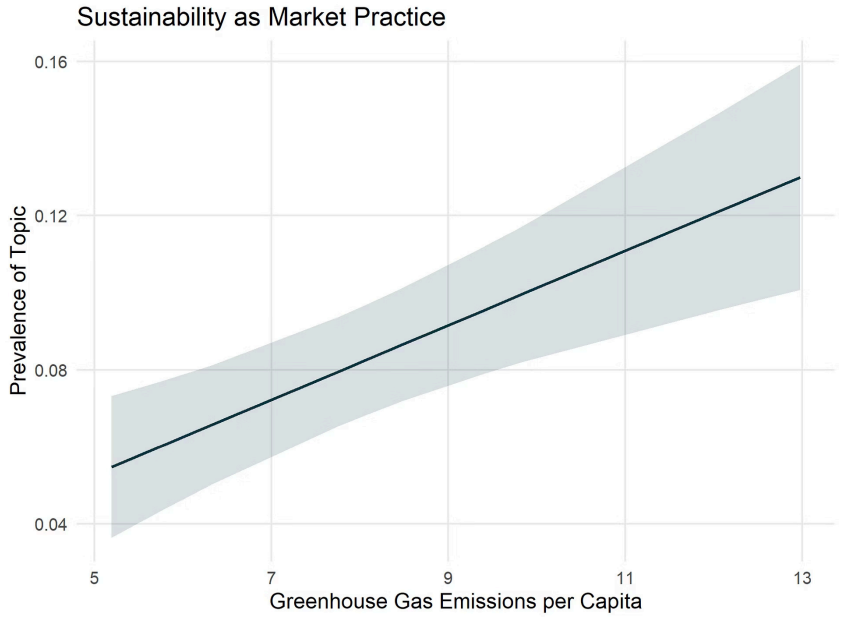
Model Specification

OLS regression models with variance inflation factor assessment for multicollinearity.

Emissions and Sustainability Discourse

These findings suggest **asset owners from more carbon-intensive economies are more likely to adopt these sustainability topics** in annual reporting, potentially reflecting regulatory pressures or stakeholder expectations.





National Context Matters

Netherlands: Emissions Focus

Dutch pension funds emphasise CO_2 emissions measurement and reduction targets, driven by the Climate Commitment requiring annual portfolio emissions reporting.

Cross-European: Market Integration

Asset owners from various countries adopt sustainability as market practice discourse, integrating environmental concerns with financial considerations.

Sweden: Leadership Ambitions

Swedish asset owners present themselves as sustainability leaders with net zero commitments, operating within supportive regulatory environment including Climate Act.

The VIDI Project: A Hybrid Design

The VIDI project employs a sophisticated methodological framework, integrating multiple generations of text analysis techniques to achieve a coherent and robust research design.

Exploratory Layer: Topic Modelling

Identifies latent themes and patterns in large datasets, offering an initial, unsupervised understanding of the discourse structure.

Sample Expansion

Broaden analysis beyond top 100 to include **400 largest asset owners globally**, enabling comprehensive evaluation across diverse firm sizes and institutional contexts.

Document Diversity

Expand corpus to include sustainability reports, SFDR disclosures and climate change reports alongside annual reports.

Analytical Layer: RAG and Multi-agent LLM

Conducts in-depth analysis and classification, leveraging advanced LLM capabilities for enhanced accuracy and nuanced interpretation.

The Corpus: Global Asset Owners

We collected annual reports from the **400 largest asset owners globally**, covering pension funds, insurance companies and sovereign wealth funds.

~40K	+12M	48	13
Pages	Tokens	Countries	Languages
Total document pages	Textual data for processing	Worldwide coverage	Mainly English

Conceptions of Green and Sustainable Finance

1

Sustainable Development

Investment in sustainable development, including references to the United Nations Sustainable Development Goals (SDGs).

2

Responsible Investment (ESG)

Financial activities that emphasise the consideration of environmental, social and governance factors (ESG) and that may include references to active ownership, ESG integration or risk analysis.

3

Green Growth

Financial investment that emphasises the long-term for the realisation of economic growth alongside environmental sustainability.

4

Net Zero

Aligning financial activities with the Paris Agreement goals, achieving net-zero carbon emissions by 2050.

5

Decarbonization

Financial activities that help reduce or eliminate carbon emissions, including references to emissions targets or GHG emissions.

6

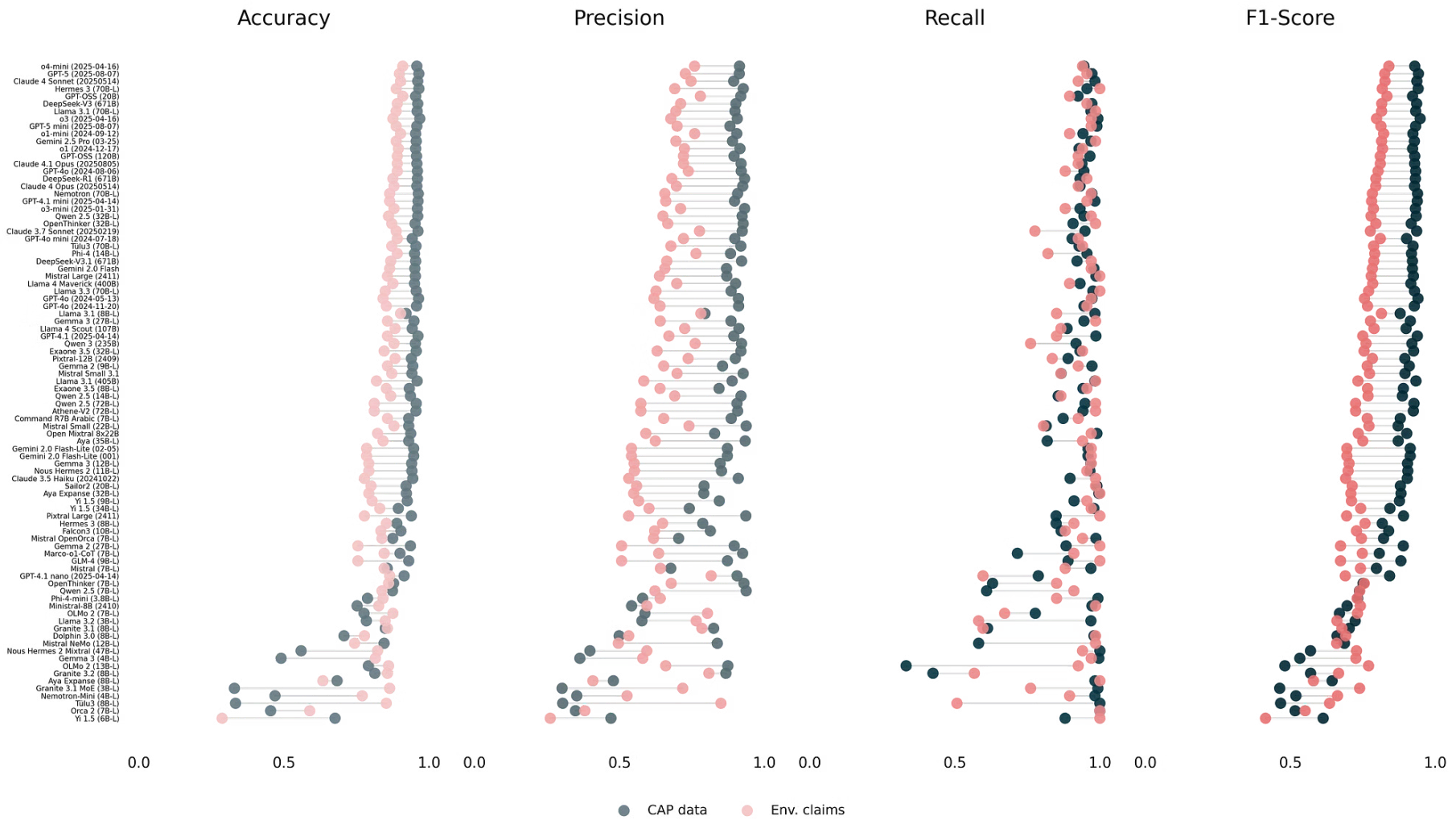
Transition Finance

Financial activities targeted at sectors or companies with high carbon emissions and to reduce their negative impact on the environment and society.

7

Conservation Finance

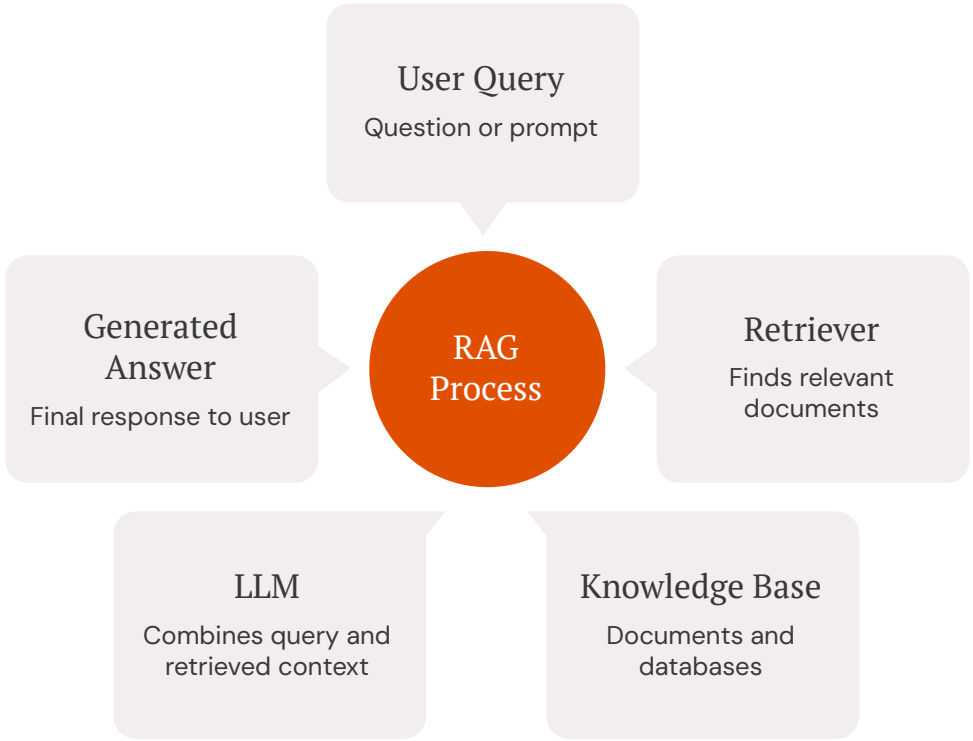
Financial activities to preserve or repair nature (land, water, natural resources).

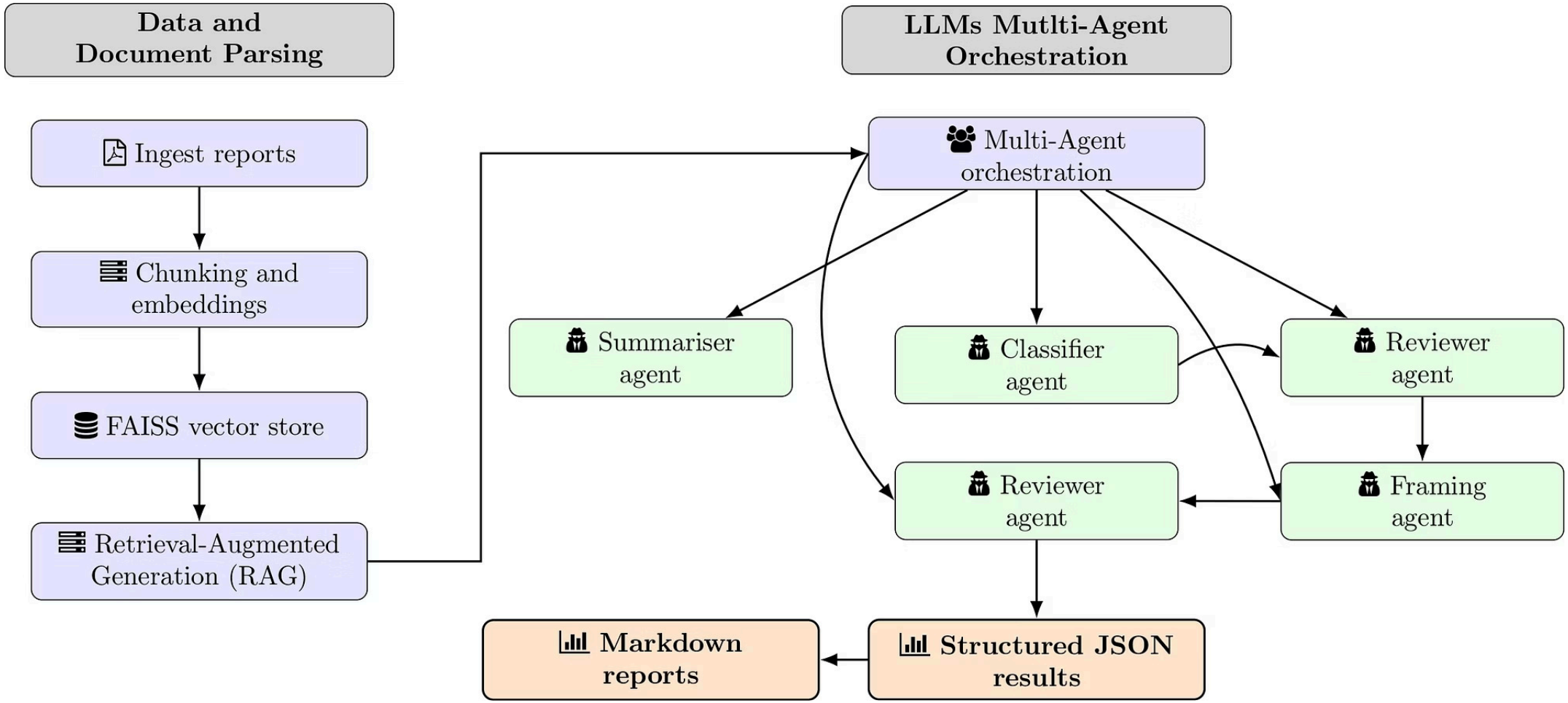

Accuracy
Precision
Recall
F1-Score
o4-mini (2025-04-16)
GPT-5 (2025-08-07)
Claude 4 Sonnet (20250514)
Hermes 3 (70B-L)
GPT-OSS (20B)
DeepSeek-V3 (671B)
Llama 3.1 (70B-L)
o3 (2025-04-16)
GPT-5 mini (2025-08-07)
o1-mini (2024-09-12)
Gemini 2.5 Pro (03-25)
o1 (2024-12-17)
GPT-OSS (120B)
Claude 4.1 Opus (20250805)
GPT-4o (2024-08-06)
DeepSeek-R1 (671B)
Claude 4 Opus (20250514)
Nemotron (70B-L)
GPT-4.1 mini (2025-04-14)
o3-mini (2025-01-31)
Qwen 2.5 (32B-L)
OpenThinker (32B-L)
Claude 3.7 Sonnet (20250219)
GPT-4o mini (2024-07-18)
Tülu3 (70B-L)
Phi-4 (14B-L)
DeepSeek-V3.1 (671B)
Gemini 2.0 Flash
Mistral Large (2411)
Llama 4 Maverick (400B)
Llama 3.3 (70B-L)
GPT-4o (2024-05-13)
GPT-4o (2024-11-20)
Llama 3.1 (8B-L)
Gemma 3 (27B-L)
Llama 4 Scout (107B)
GPT-4.1 (2025-04-14)
Qwen 3 (235B)
Exaone 3.5 (32B-L)
Pixtral-12B (2409)
Gemma 2 (9B-L)
Mistral Small 3.1
Llama 3.1 (405B)
Exaone 3.5 (8B-L)
Qwen 2.5 (14B-L)
Qwen 2.5 (72B-L)
Athene-V2 (72B-L)
Command R7B Arabic (7B-L)
Mistral Small (22B-L)
Open Mixtral 8x22B
Aya (35B-L)
Gemini 2.0 Flash-Lite (02-05)
Gemini 2.0 Flash-Lite (001)
Gemma 3 (12B-L)
Nous Hermes 2 (11B-L)
Claude 3.5 Haiku (20241022)
Sailor2 (20B-L)
Aya Expanse (32B-L)
Yi 1.5 (9B-L)
Yi 1.5 (34B-L)
Pixtral Large (2411)
Hermes 3 (8B-L)
Falcon3 (10B-L)
Mistral OpenOrca (7B-L)
Gemma 2 (27B-L)
Marco-o1-CoT (7B-L)
GLM-4 (9B-L)
Mistral (7B-L)
GPT-4.1 nano (2025-04-14)
OpenThinker (7B-L)
Qwen 2.5 (7B-L)
Phi-4-mini (3.8B-L)
Ministral-8B (2410)
OLMo 2 (7B-L)
Llama 3.2 (3B-L)
Granite 3.1 (8B-L)
Dolphin 3.0 (8B-L)
Mistral NeMo (12B-L)
Nous Hermes 2 Mixtral (47B-L)
Gemma 3 (4B-L)
OLMo 2 (13B-L)
Granite 3.2 (8B-L)
Aya Expanse (8B-L)
Granite 3.1 MoE (3B-L)
Nemotron-Mini (4B-L)
Tülu3 (8B-L)
Orca 2 (7B-L)
Yi 1.5 (6B-L)
0.0
0.5
1.0
CAP data
Env. claims

Retrieval-Augmented Generation (RAG)

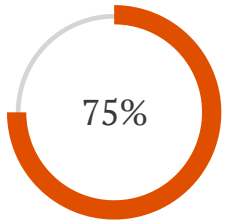
Retrieval-Augmented Generation (RAG) is a powerful technique that enhances LLMs by providing them with access to external and domain-specific information.

This approach significantly mitigates issues such as knowledge cut-offs and factual inaccuracies (hallucinations).

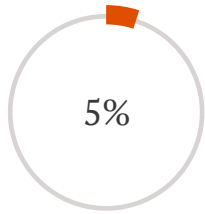



Data and Document Parsing
Ingest reports
Chunking and embeddings
FAISS vector store
Retrieval-Augmented Generation (RAG)
LLMs Mutlti-Agent Orchestration
Multi-Agent orchestration
Summariser agent
Classifier agent
Reviewer agent
Reviewer agent
Framing agent
Markdown reports
Structured JSON results

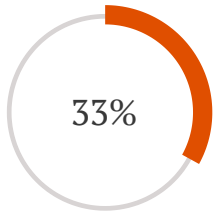
Batch-01



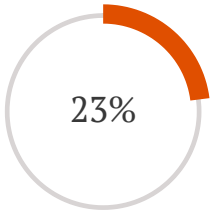
Batch-02



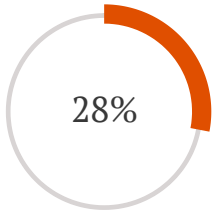
Batch-03



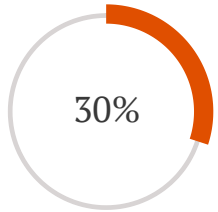
Batch-04



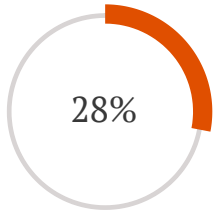
Batch-05



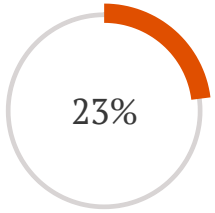
Batch-06



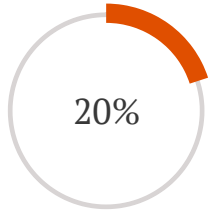
Batch-07



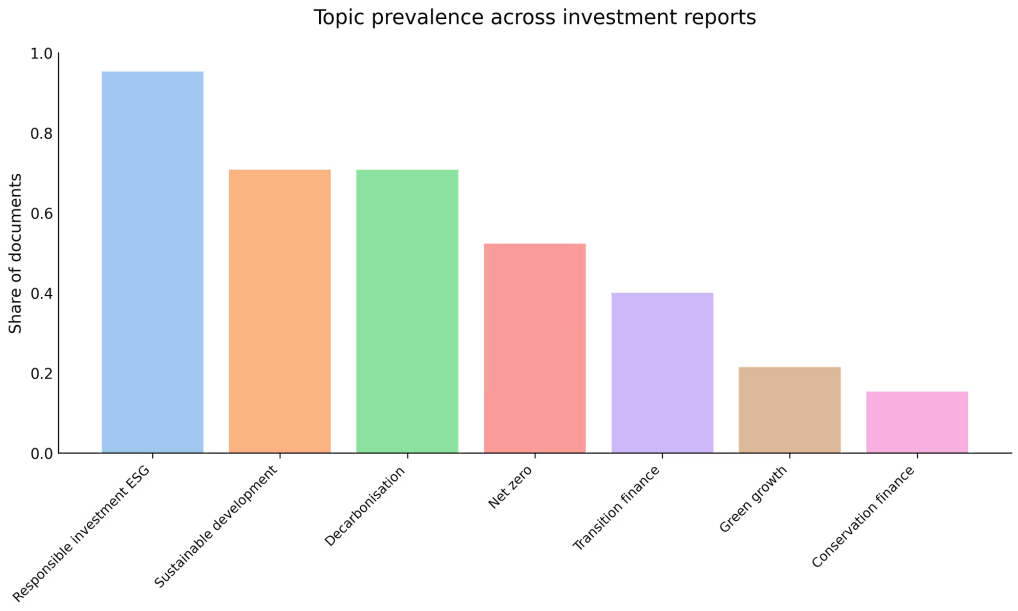
Batch-08

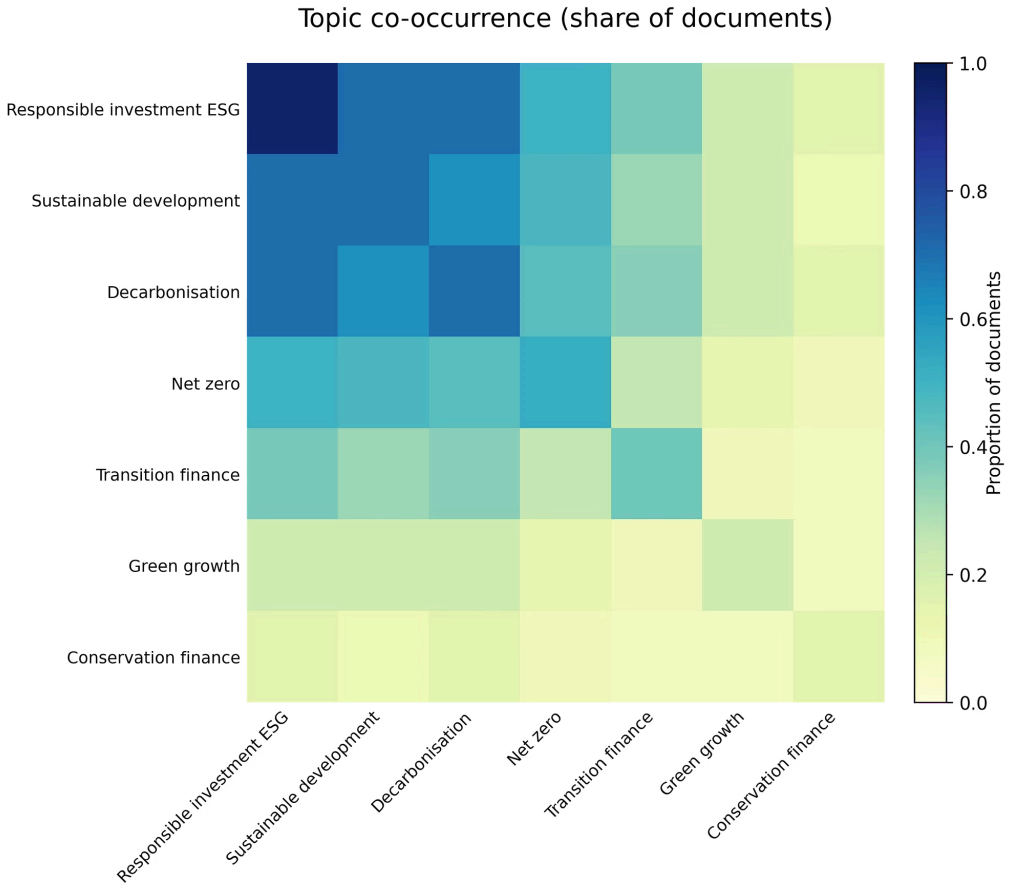


Batch-09



Batch-10


Topic prevalence across investment reports
Share of documents
1.0
0.8
0.6
0.4
0.2
0.0
Responsible investment ESG
Sustainable development
Decarbonisation
Net zero
Transition finance
Green growth
Conservation finance


Topic co-occurrence (share of documents)
Responsible investment ESG
Sustainable development
Decarbonisation
Net zero
Transition finance
Green growth
Conservation finance
Proportion of documents
1.0
0.8
0.6
0.4
0.2
0.0

Substantive Implications for CPE

The application demonstrates that LLMs can extract meaningful signals from highly heterogeneous annual reports.

Cross-National Mapping

Create comprehensive map of sustainability discourse linked to institutional variables such as varieties of capitalism typologies and regulatory frameworks.

Institutional Testing

Test whether coordinated market economies privilege patient green capital or whether liberal systems compensate through market innovation.

Reproducibility

Deterministic decoding based on local deployment and open-source LLMs (**GPT-OSS, Llama, Hermes-3**). Possibility of swapping models with minimal performance loss.

Thank you very much!

bgonzalezbustamante.com